# AGGREGATING ORGANIZATIONAL PERFORMANCE METRICS USING THE ANALYTIC HIERARCHY PROCESS

THESIS

Bret L. Indermill, Captain, USAF

AFIT/GSM/LAS/95S-3

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GSM/LAS/95S-3



# AGGREGATING ORGANIZATIONAL PERFORMANCE METRICS USING THE ANALYTIC HIERARCHY PROCESS

THESIS


Bret L. Indermill, Captain, USAF



AFIT/GSM/LAS/95S-3


Approved for public release; distribution unlimited

AFIT/GSM/LAS/95S-3

# AGGREGATING ORGANIZATIONAL PERFORMANCE METRICS USING THE ANALYTIC HIERARCHY PROCESS

THESIS

Presented to the Faculty of the School of

Logistics and Acquisition Management

Air Education and Training Command

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Systems Management

Bret L. Indermill, B.S.

Captain, USAF

## Preface

This research explored the feasibility and desirability of applying the Analytic Hierarchy Process (AHP) to the task of scoring Quality Air Force (QAF) unit self-assessment (USA) reports. Unfortunately, the proposed AHP-USA scoring method did not prove to be significantly better than the current QAF-USA method.

This research was undertaken because a general need was expressed for methods to combine lower-level performance metrics into higher-level aggregate scores or performance indices. The QAF-USA process was targeted because it had proved to be a relatively inconsistent and time-consuming approach for scoring USA reports.

There are many people that helped me survive this thesis effort. First, I thank my advisors, Major Cindy Fossum and Lt. Col. David Murphy, for their support and guidance throughout my research, writing, and editing. I also wish to thank Captains Ken Hamner and Ken Kessler for their sponsorship, guidance, and introduction to Ms. Sande Staub, from whom I received truly astounding support; without her help I would still be trying to collect my data. Also, assignments are made or broken by the people with whom you work and play. AFIT is no exception; thus, I largely owe my sanity and success to the entire GSM-95S gang, and to Mari and Smitty in particular. Also, I am very grateful for the love and friendship of Catherine Warrick. Her encouragement came regularly via e-mail. Finally, I would like to thank my parents, Roy and Kathryn, for their love, their sacrifices, and their devotion to providing a good, happy home for all their children.

Bret L. Indermill

## Table of Contents

## List of Figures

## List of Tables

Table | Page

## List of Equations

AFIT/GSM/LAS/95S-3


## Abstract

Measurement provides factual information which is necessary for effective control of business processes. Implementing a Total Quality Management (TQM) philosophy is a common process in many organizations today. The United States Air Force is using the Malcolm Baldrige National Quality Award criteria to measure organizational performance in implementing the Quality Air Force (QAF) initiative. Unfortunately, the Baldrige-based unit self-assessment (USA) process is an inconsistent measure due to its subjectivity, and is also time consuming to use.

This study determined that a new USA method based on the analytic hierarchy process (AHP) was not a significant improvement over the existing QAF-USA method. Specifically, this study developed a new USA scoring method by adapting the AHP to use existing QAF evaluation criteria. A group of 11 evaluators used this new AHP-USA method to score a portion of a USA report, and they also compared the AHP-USA method to the QAF-USA method to gauge its understandability, usability, believability and applicability. The resulting data was used to determine the overall feasibility and desirability of using the new method as a replacement for the QAF-USA method.

AGGREGATING ORGANIZATIONAL PERFORMANCE METRICS

USING THE ANALYTIC HIERARCHY PROCESS

# I. Introduction

## Research Objective

This research involves a new application of the analytic hierarchy process (AHP) to explore the feasibility and desirability of using the AHP to aggregate organizational performance metrics. The AHP is a measurement theory and multicriterion decision-making process developed by Thomas L. Saaty at the Wharton School of the University of Pennsylvania from 1971 to 1975 (Saaty, 1987:161). Performance metrics are commonly used to assess the progress of activities (e.g., programs, projects, tasks) that an organization undertakes, thus giving insight into the performance of the organization. An aggregate measure formed by combining a variety of performance metrics can provide insight into the overall performance of the organization or one of its divisions. In turn, such insight may help build the "profound knowledge" which is a key element of Total Quality Management (TQM) based on the teachings of W. Edwards Deming (Aguayo, 1990:49). Unfortunately, it can be very difficult to combine a diverse set of metrics into a single aggregate measure which will adequately show actual performance relative to a desired performance baseline.

## Effective Control Requires Good Data

Control is one of the fundamental functions of management. To effectively control a project, a manager must make sound decisions based on accurate, timely data. Often, this data can be organized into three dimensions of project performance: cost, schedule, and technical (Nicholas, 1990:22). If the *project* being managed is relatively small or simple, then gathering and organizing the requisite performance data should also be relatively simple. Furthermore, it should be relatively easy for a manager to interpret this small body of data. However, when a project becomes much more complex, when several projects make up a larger *program*, or when a program must be measured by more subjective criteria, it becomes increasingly difficult to take all the pertinent data into consideration when making decisions.

In such cases, an aggregate measure, which combines many separate criteria into a single score or index, is often needed to adequately capture a complex process. For instance, implementing Total Quality Management (TQM) practices within the United States Air Force (USAF) is just such a process. To enable a unit to gauge its progress towards TQM principles, the Air Force created a unit self-assessment (USA) process based on the evaluation criteria used for scoring applications for the Malcolm Baldrige National Quality Award. This USA process is also used to determine the Secretary of the Air Force Unit Quality Award Winner. Furthermore, the evaluation criteria used in the USA process also serve other purposes as outlined by the Quality Air Force (QAF) Criteria booklet (AFQI, 1993:1); specifically, the criteria are intended to:

- serve as a working tool for planning, assessment, training, and other uses;
- raise quality performance expectations and standards;
- facilitate communication and sharing among and within organizations of all types based upon a common understanding of key quality and operational performance requirements.

## The Problem

Unfortunately, a couple of weaknesses in the QAF-USA method have been experienced by units trying to use it to assess their TQM performance. Specifically, the Aeronautical Systems Center (ASC) at Wright-Patterson Air Force Base (WPAFB), Ohio, has discovered that the Baldrige-based method suffers from two problems:

1. It is inconsistent, due to its subjectivity. Specifically, USA scores for the same unit can vary significantly (over 30 percentage points) between evaluators.

2. It can be very time consuming. Thousands of hours have been spent preparing and scoring USA reports.

Thus, there is interest in exploring other methods which have the potential to improve scoring consistency while reducing the time and effort required to perform the unit self-assessment. With these objectives in mind, and an eye towards the broader function of aggregating metrics, this study set out to find a better USA scoring method.

## Methodology

In order to gain a better perspective of the problem and potential solutions, this study explored the literature about two basic topics before developing and testing a new USA scoring method. First, it looked at the basics of *performance measurement* and *aggregation*. This examination included compiling a list of methods which have the potential to aggregate organization performance metrics. Then, one of these methods (the AHP) was selected for a detailed evaluation.

Figure 1 summarizes the process used to evaluate the proposed AHP-based USA scoring method. As shown, a simple experiment was performed to compare the AHP-USA scoring method with the existing Quality Air Force unit self-assessment (QAF-USA) scoring method. A group of eleven USA evaluators scored a small portion of a recent USA report using both scoring methods. The results from these two scorings, along with responses to a group of feedback questions, were then analyzed to determine whether the AHP-USA method was both feasible and desirable as a replacement for the current QAF-USA scoring method.

## Overview

The remainder of this report presents the research, data, analysis, conclusions, and recommendations of this study. This information is organized into four chapters outlined here.

Chapter II: Literature Review. This chapter provides background information gleaned from a variety of sources. It also provides a much more detailed treatment of the objectives of this research. Specifically, the following topics are addressed:

- Performance measurement
- Aggregation
- The analytic hierarchy process
- The unit self-assessment process
- Research questions

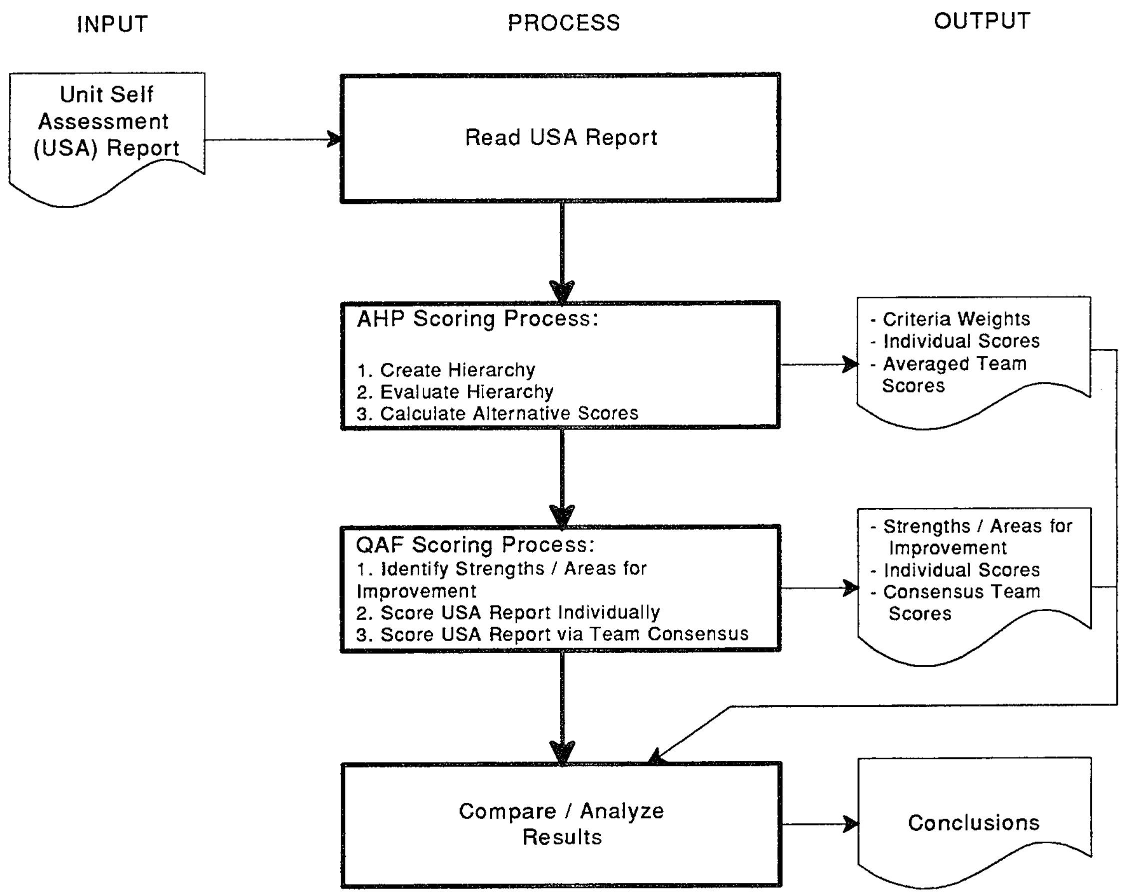



**Figure 1 Study Methodology**

Chapter III: Methodology. This chapter expands on the study methodology just outlined. Specifically, it discusses the inputs, processes, and outputs shown in Figure 1. This discussion includes a description of the data collection instrument and approach. Also, pre- and post-experiments steps used to implement the AHP-USA method are described. Finally, the chapter covers the quantitative and qualitative methods used to analyze the data.

Chapter IV: Data Description, Analysis, and Findings. This chapter presents all of the data collected during the study, the analysis of that data, and the findings arising from the analysis. In particular, this chapter holds the answers to the research questions.

Chapter V: Conclusions and Recommendations. This chapter summarizes the findings of Chapter IV and draws more general conclusions regarding the fundamental questions of feasibility and desirability. It also offers recommendations for potential follow-on investigations into applications of the AHP.

# II. Literature Review

## Introduction

This chapter defines performance measurement including measurement strategies, discusses reasons for measuring performance, identifies the need for aggregation, and discusses the capabilities of five potential aggregation methods. Also, a detailed description of the QAF-USA scoring method provides a foundation for better understanding the proposed AHP-USA scoring method. Finally, the detailed research questions are presented.

## Need for Performance Measurement

Aeronautical Systems Center (1990:9) expressed its position on the value of measurement as follows: "Measurement is the springboard to improvement and allows the organization to take quality out of the subjective by quantifying conformance to requirements." For example, consider a system program office (SPO) which is building a replacement satellite command and control system. Assume the main objective is well defined: *improve maintainability and reduce operating costs by replacing the existing mainframe computer hardware and software with Ada code developed for modern workstations and personal computers*. In addition to this main objective, there are many subordinate objectives that need to be met for success. Specifically, the program manager might have to stay within a 15 million dollar *cost* constraint, complete the project within a 20 month *schedule*, and satisfy all technical *performance* requirements (e.g., command rate to the spacecraft, reliability, availability, and maintainability). Given that all these objectives are desirable, how will measurement help the organization achieve them?

Measurement aids progress towards goals and objectives by fulfilling two basic needs. First, measurement provides a frame of reference for assessing progress. A frame of reference is nothing more than a measurement scale which helps people assess absolute progress (e.g., total costs to date) or relative progress (e.g., percent of software modules tested this month compared to last month). Furthermore, the frame of reference does not have to be internal to the organization. Instead, another organization can be used to set the standard (benchmark) for a particular measure or a pre-defined scale can be used to assess progress.

Second, measurement can help bridge the gap between managers and workers by giving both a common, objective perspective. In essence, measurement gives people information which is necessary to make rational operational decisions (Brinkerhoff and Dressler, 1990:9,22). Clearly, performance measurement is necessary to provide information for rational operational control. With this understanding of the need for measurement, we take a look at how performance measurement is defined.

## Definition of Performance Measurement

The Air Force Quality Center (1993:Glossary) defines a metric as:

> A measurement, taken over a period of time, that communicates vital information about a process or activity. A metric should drive appropriate leadership or management action. Physically, a metric package consists of an operational definition, measurement over time, and presentation.

This description generically defines a metric as a *measurement* that communicates information about a *process* or *activity*. But, what, exactly, is being measured? In short: performance.

Brinkerhoff and Dressler (1990:16) define performance measurement as a superset of productivity measurement. Performance measurements can be designed to measure any aspect of an organization, while productivity measurements are more strictly defined as ratios of output to input, such as bushels of wheat harvested per gallon of diesel. At its roots, performance measurement is like any other measurement process. Specifically, it involves assigning a symbol (e.g., a number) to an observable organizational attribute or event (Cooper and Emory, 1995:141). Ideally, this assignment (mapping) process is purely *objective* in that the attributes or events (e.g., bushels of wheat harvested) are easily observed and can be unambiguously mapped to a symbol (e.g., 1200). However, in some cases the mapping is more *subjective* because the attributes or events are not directly observable or the mapping rules themselves leave room for interpretation. In short, performance measurements may be either objective or subjective. Similarly, productivity measurements can also be objective or subjective, but they are also classified by whether they show a snapshot of productivity at a point in time or the change in productivity over a period of time.

Specifically, Sink (1985:25) defines these two types of productivity measures as *static* and *dynamic*. A static measure is a single output/input productivity ratio (e.g., bushels of wheat / gallons of diesel) taken at a point in time. A dynamic measure, on the other hand, is the ratio of two static measures taken at two different times:

$$\frac{(\text{Bushels of Wheat/Gallons of Diesel})_{\text{Week 2}}}{(\text{Bushels of Wheat/Gallons of Diesel})_{\text{Week 1}}} \tag{1}$$

This results in a dimensionless measure which shows change from the previous period. For example, if 100 bushels were harvested for every gallon in the first week and 125 bushels per gallon were harvested in the second week, then the dynamic ratio would be 125/100 or 1.25. The measure is *dynamic* because it only shows that productivity has *increased* by 25 percent over the previous week, it does not reflect the *amount* of wheat harvested per gallon.

Sink also defines three types of measures that are classified by the number of input *classes* that are included in the measure. Table 1 summarizes these definitions:

TABLE 1

MEASUREMENT DEFINITIONS

| Type of Measure | Number of Input Classes in the Denominator | Measure Examples $\left(\frac{\text{Generic Output}}{\text{Input Class(es)}}\right)$ |
|---|---|---|
| Partial-factor | 1 | $\frac{Widgets}{Labor}$ |
| Multifactor | > 1<br>< All | $\frac{Widgets}{(Labor + Materials)}$ |
| Total-factor | All Classes | $\frac{Widgets}{Labor + Capital + Energy + Data + Materials}$ |

(Sink, 1985:25-26)

As shown, a partial-factor measure only uses a single type of input in the denominator. For example, labor or capital, but not both. A multifactor measure can include two or more types of input measures. The advantage of a multifactor measure is that it gives a broader, more representative, measure of productivity; in essence, it is a simple *aggregate* measure. Finally, a total-factor measure uses all of the five input classes listed in Table 1. For example, a total-factor productivity measure for an automobile manufacturer might be defined by taking the total number of cars produced and dividing by the costs of the assembly line workers, depreciation on the factory, electricity to run the line, development of the data to run the numerical controlled milling machines, and the steel used in the body of the vehicle. According to Sink (1985:26), the definition hinges on the number of *classes* in the denominator, not the number of individual measures. In other words, a total-factor measure does *not* have to measure *every* input into a process, it only has to include a single measure from each of the five categories shown in Table 1.

Downs and Larkey (1986:8-9) do not make this distinction. Instead, they define multifactor or total-factor measures based on the number of individual inputs in the denominator. Therefore, a total-factor measure would include all applicable inputs, not just one of each input class. Thus, a total-factor measure provides a more accurate representation of overall performance than a multifactor measure, but at the cost of collecting and processing additional data. With this clearer understanding of what constitutes performance measurement, it is time to turn to the types of measurement strategies available to an organization. Specifically, this involves an examination of three broad strategies that an organization can employ to define and structure performance and productivity measurements.

## Performance Measurement Strategies

An organization can design a performance measurement system using two basic strategies and a third hybrid approach. The first strategy is centralized. It develops a group of standard metrics by working from the top of the organization down to the division and department levels. The use of these standard metrics is mandated throughout the organization. Because this strategy gathers data from all divisions of the organization using the same measures, the measures can be relatively easily combined into a performance rating for the organization as a whole. The second strategy is decentralized. It allows work groups to develop measures which are best suited for their operations. However, the disparity of measures between groups can make it very difficult to combine the measurements for higher level management. The third hybrid approach combines the first two by establishing a minimal set of standardized measures while still permitting the work groups to develop unique measures (Sink, 1985:73).

When developing a performance measurement strategy, it is important to understand that organizational performance is inherently multidimensional. Performance is multidimensional because organizations are inherently designed to divide large, complex tasks into smaller, specialized pieces (dimensions) that can be accomplished by individuals. For example, a baseball team is an organization comprised of managers, hitters, outfielders, infielders, pitchers and catchers. Each functional group plays a role in the overall performance of the organization. Clearly, it would be inaccurate to measure the performance of the team by focusing on a single function, such as *pitching*, and therefore gauge organizational performance based on single measure, such as *strikeouts per game.*

Instead, many other performance measures are needed to get a better picture of team performance. Some examples are: dropped balls (errors) per game, hits per game, runs scored per game, bases stolen per game. Each of these measure a different dimension of the organization and therefore help form a more representative performance measure. The importance of this principle is driven home by Provost and Leddick (1993:477) by posing a hypothetical question: Would you be comfortable if your doctor relied solely on body temperature to diagnose an illness? Provost and Leddick predict a typical reader's response to the question:

> "Ridiculous!" you say. "Everybody knows that you have to look at the health of the patient from other perspectives than just temperature. You'd have to look at blood pressure, too. And heart rate. And maybe even brain functioning and reflexes. It takes more than temperature to diagnose and monitor a patient's health. You might even have to look at how some of these measures relate to each other, like heart rate and blood pressure. Look at only one? You've got to be kidding!"

When put in these terms, the importance of assessing organizational performance in a comprehensive manner using multiple measures becomes a matter of common sense.

### Organizational Performance Metrics and Aggregation

As we have seen, an organizational performance metric is a measure of an attribute or activity of an organization. A measure can be objective (e.g., the number of candles produced in a day) or subjective (e.g., the motivation level of member of the candle makers union). It can also be dynamic (e.g., candle orders show a 15 percent decline from last year) or static (e.g., 12,300 birthday candles were sold last year). Such performance measures can apply to organizations which range in size from a small work team to multi-national corporations. Regardless of the size of the organization, a variety of metrics can be used to help measure organizational performance.

However, each metric will have a different importance depending on the users' needs and their positions in the organization. This difference will tend to widen as the size of the organization increases. For example, a measure of the failure interval for a particular milling machine would have significance for the maintenance person who has to keep it running. However, the same measure would have little significance to the president of the corporation. But, if many such measures, from many different machines, could be collected and combined into a single *failure rate* metric, then this might be a measure of the health of the company's infrastructure or maintenance procedures as a whole. This is one example of an aggregate measure. If a meaningful way can be found to combine metrics that measure a much wider variety of attributes, then it should be possible to create a few, or even a single, aggregate metric for an entire organization.

### Need for Aggregation

As discussed in Chapter I, the Aeronautical Systems Center (ASC) is using the Quality Air Force's unit self-assessment (QAF-USA) process to help gauge progress towards TQM principles. This process evaluates unit performance by measuring 28 specific *items* -- which fall into 7 broader functional *categories* -- and then aggregating the scores for each of the *items* into an overall performance score. (A more comprehensive description of the QAF-USA method is provided later in this chapter.) Here then, is an example of a far-reaching, complex process (implementing TQM) which requires a comprehensive multidimensional measure to assess progress. Granted, the overall score provided by the QAF-USA method, like any aggregate measure, may hide the specific cause(s) for a particular performance level, and thus in the USA case is probably not as valuable as the narrative feedback generated during the evaluation (scoring) process. However, this performance score does provide a snapshot of performance, and is used to help identify candidates for the Secretary of the Air Force Quality Award (AFQI, 1993:1).

Aggregating other metrics (e.g., cost, schedule, and performance measures) into an overall performance index could also have value by providing top level ASC managers with a means to monitor performance at an organizational level and flag any undesirable trends. For example, the Air Force Material Command's Technology Transfer Office is searching for an effective way to measure their performance at working levels within the organization and then package this data so that it can be used by middle and upper management. Each management layer in the organization needs a different level of detail, implying that a method to aggregate detailed, low-level metrics into less detailed, higher-level metrics might meet their needs (Guilfoos, 1994:2). Clearly there is a need for an effective aggregation method; unfortunately, it can be a challenging task to aggregate a variety of performance metrics.

## Difficulty of Aggregation

Meaningfully combining different types of performance measures into a single index is a major problem (Sink, 1985:32). Consider the baseball team measures. Some measures are *good* if they are close to zero (e.g., errors per game) while others (e.g., number of hits per game) are good if they are high. Similar relationships occur in business (e.g., inventory levels and gross sales). Even if two measures are on roughly the same *goodness* scale (e.g., number of hits per game and number of stolen bases per game), are they *equal* in contribution to the team's goal of winning? Are *number of satisfied workers* and *net profits* of equal importance when measuring organizational performance? Such complexities make it difficult to develop a single performance measure (Downs and Larkey, 1986:91). However, methods exist which are either specifically designed to aggregate data or have the potential to do so. Before examining these methods, a brief overview of a measurement and aggregation process is needed.

## Definition of Aggregation

Generically, aggregation is the act of combining several parts into a whole. In the context of measurement, it means to combine several metrics into a single metric. Figure 2 provides an overview of how this process works.

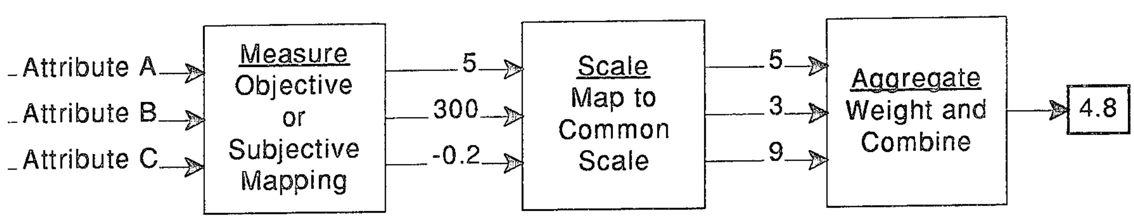



Figure 2 Aggregation Process

As shown, the process begins with measurement. This involves assigning a number (or other meaningful symbol) to each observed or estimated state of an attribute. Ideally, this mapping is done objectively; however, sometimes an attribute cannot be directly and objectively observed. In these cases, a subjective judgment is required to perform the mapping. Next, this raw measurement is scaled by another mapping process which translates the raw measurement from its inherent or fundamental scale to a common scale which represents value to the user. For example, the speed of an aircraft may be objectively measured as 300 miles per hour. This speed then translates to a value of 3 on a common 1 to 10 scale, perhaps because 300 mph is only 30 percent of the user's desired speed. It is possible that these first two steps could be combined; however, for our purposes, they will be treated separately. Finally, the individual values from the common scale are aggregated. Typically, this is done by weighting the individual values based on their perceived importance and then adding them together to calculate the overall score or index. The aggregation model described here provides a framework for the following discussions and evaluations of different aggregation methods. So, with this framework in hand, we proceed to consider several different methods which may be applicable to the USA scoring process.

## Aggregation Methods

A survey of performance measurement literature revealed five methods which, upon a cursory review, appeared to be the most reasonable candidates to replace the aggregation method used in the QAF-USA scoring process. This section provides the results of a more detailed analysis of the five potential replacements and the QAF-USA scoring method. The intent of this analysis was to select one of the five aggregation methods for use in this research. Specifically, the six methods discussed here are:

1. The Normative Productivity Measurement Methodology (NPMM)
2. The Multifactor Performance Measurement Model (MFPMM)
3. The Multicriteria Performance/Productivity Measurement Technique (MCP/PMT)
4. Data Envelopment Analysis (DEA)
5. The Analytic Hierarchy Process (AHP)
6. The Quality Air Force Unit Self-assessment (QAF-USA) Process

Some of these methods have been specifically designed to aggregate metrics, while others are more general decision tools that have their roots in group dynamics, operations research, or multiple criterion decision making (MCDM).

Before discussing each of these methods, the decision criteria which were used to select an aggregation method (from the five potential replacement methods) for use in this study are presented. Then, brief descriptions of the methods and their evaluations follow.

The decision criteria can be grouped into two classes: general and specific. The general criteria are derived from desirable characteristics of any aggregation method, while the specific criteria focus on desirable attributes for the USA application. The criteria in each class are described below.

<u>General Evaluation Criteria</u>. A good general aggregation method should be mathematically correct, and usable in a wide variety of measurement environments and by a wide variety of people. In other words, a good method should be *logical, adaptable* and *simple*. By referring back to the three basic steps of the aggregation model in Figure 2,

these general criteria can be more specifically defined. First, a logical aggregation method will not violate mathematical principles involved with manipulating nominal, ordinal, interval, and ratio scales. For purposes of evaluating aggregation methods, a rigorous mathematical evaluation was not used. Instead, methods were evaluated based on a subjective assessment of how well each step in the aggregation model was defined within the context of the method. The rationale being that a well defined method is less likely to be misused (i.e., use in an illogical manner). Second, an adaptable aggregation method will provide for measurement of a wide variety of attributes. This means accommodating both objective and subjective measurements, and various measurement scales. Also, an adaptable method should provide for input from the user with regard to value mappings and weightings. Finally, an aggregation method should be simple to use. This means that the measurement, scaling, and aggregation steps should all be easy to understand and implement.

Specific Evaluation Criteria. For the purposes of finding a replacement for the QAF-USA process, three specific criteria were used to evaluate aggregation methods which might be incorporated into a new scoring method. First, any replacement scoring method must be able to adapt to the existing QAF evaluation criteria. This means that the *attributes* which will be measured are already defined by the QAF *criteria*. This probably also means that the basic subjective approach used to measure unit progress based on these criteria will need to be adopted from the QAF-USA method. Second, the new scoring method should be able to produce a score on the same 0 to 100 percent scale as the current QAF-USA method. Third, the new scoring method should have a reasonable potential to be an improvement (i.e., more accurate and/or efficient) over the existing QAF-USA method.

Table 2 summarizes the evaluation results for each aggregation method, based on the above criteria. The discussion that follows the table elaborates on these results.

TABLE 2

AGGREGATION METHOD SELECTION

| | NPMM | MFPMM | MCP/PMT | DEA | AHP |
|---|---|---|---|---|---|
| Logical (well defined) | + | ++ | ++ | ++ | ++ |
| Adaptability | ++ | - | ++ | 0 | ++ |
| Simplicity | ++ | -- | ++ | - | + |
| Use QAF Criteria | ++ | -- | ++ | -- | ++ |
| Produce USA Score | ? | -- | 0 | - | 0 |
| Improvement Potential | 0 | -- | 0 | -- | + |

<u>Normative Productivity Measurement Methodology</u>. The NPMM is fairly adaptable and simple to implement, yet is less desirable than two other methods for two key reasons. First, it does not define how performance metrics should be aggregated, thus making it impossible to assess its potential for producing a suitable USA score. Its lack of definition is primarily due to the fact that the NPMM is designed to help an organization *develop* metrics using the nominal group technique (NGT), not *aggregate* them. However, the NPMM, might still be used as a simple aggregation tool by weighting the performance metrics based on the rank assigned to each metric during one of the steps of the NGT. An aggregate metric could then be calculated by taking the weighted sum of the metrics. In fact, the MCP/PMT method extends the basic NPMM, albeit more elaborately

than the simple weighted sum concept presented here. The second drawback of the NPMM is that it was judged to have little if any improvement potential over the existing QAF-USA method and significantly less potential than other methods. This assessment was based on the understanding that the NGT can be a long and tedious process which has the potential to require excessive man-hours to implement, thus lowering its potential for economy. Of course, adaptations of the method might be able to limit repetitive use of the NGT, thus reducing this concern. However, without further work to create such an adaptation, the basic NPMM loses out to both the MCP/PMT and the AHP. (Sink, 1985:122)

Multifactor Productivity Measurement Model. The MFPMM fails on every criteria except clear logical definition. First, it is not adaptable because the MFPMM dictates the measures that will be used throughout the organization. The MFPMM is designed to measure the productivity of an entire organization based on a well defined set of measures. The organization *can* define its outputs, however, the inputs are pre-defined as labor, material, energy, investment, and services. Aggregate performance is calculated based on the dollar value of all these organizational inputs and outputs. There is no provision for weighting the importance of the inputs or outputs because the dollar value is considered a standard fixed unit of comparison. In fact, constant dollars are used throughout the model to ensure accurate historical comparisons. For example, if the manager of a tire company wanted to see how the proportion of petroleum used to make a single tire had changed over the last 10 years, MFPMM would adjust the cost of the petroleum in each of the years so that the comparisons would be made using constant (e.g., 1980) dollars. The MFPMM is essentially a performance accounting system which can be used to flag performance variances and then explore their underlying cause. Due to its ridged definition and complexity, the MFPMM receives low marks for most general criteria and all USA-specific criteria. (Sink, 1985:79-146).

Multicriteria Performance/Productivity Measurement Technique. The MCP/PMT is a strong contender among the aggregation methods considered. Like the NPMM, the MCP/PMT uses the nominal group technique to develop a set of ranked performance measures. Then, the MCP/PMT extends the NPMM process by developing a normalized rank for each performance measure. This rank is used to weight the importance of each metric. The ranking is done using pairwise comparisons where the *highest* ranked measure is used as the basis of comparison for all the lower measures. These paired comparisons, and the resulting weights, may be somewhat flawed because, according to Saaty (1993:3), the *lowest* element in a pairwise comparison serves as the most meaningful unit of comparison. Also, Wipper (1994:363), a Central Services program manager with the Oregon Department of Transportation's (ODOT's) Office of Productivity Services, writes that the process of weighting performance measures (during a recent pilot project which used an MCP/PMT-based approach) was "somewhat arbitrary." (Sink, 1985:199-204)

This does not mean that the MCP/PMT cannot be successfully applied. The ODOT project was deemed successful and Young (1992:53) also successfully used the

MCP/PMT to assign productivity scores to 22 hospitals during his research which compared aggregate ranks from the MCP/PMT to ranks from the data envelopment analysis method. Also, in many respects, the MCP/PMT is similar to the QAF-USA scoring process (which is described in detail later in this chapter). Both define how measurements will be taken and translated to a common scale. And, both provide well defined methods to aggregate the resulting measurements. Unfortunately, it was these very similarities that resulted in a neutral assessment regarding its potential to improve the QAF-USA process.

Data Envelopment Analysis. DEA is a linear programming method that can be used to calculate the relative performance of individual organizations with respect to the performance of a group of peer organizations. Although it earned high marks for logic, it is less adaptable, because DEA requires the use of productivity measures (i.e., output/input). Other effectiveness measures, such as technical performance (e.g., percent of requirements met), that do not have clearly associated inputs cannot be *directly* addressed with DEA. However, it may be possible to create an adequate efficiency metric by dividing an effectiveness output by an appropriate input measure (e.g., percent of requirements met / number of people assigned to the requirements branch). Unfortunately, this type of adaptation would tend to increase the complexity of implementing a DEA-based process. A second potential problem arises because of the approach that DEA uses to accommodate custom weights. Specifically, performance variables must be weighted *before* applying DEA. This adaptation can make it difficult to interpret the resulting aggregate measures (Epstein and Henderson, 1989:113). Therefore, it seems unlikely that the output ratings from a DEA analysis could be used to calculate a single USA-like score. Finally, considering the adaptability problems and the uncertainty about generating a suitable score, the DEA method also receives low marks for its potential to become a better scoring method than the QAF-USA method. In summary, DEA is a promising technology for comparing the relative efficiencies of many units based on a variety of efficiency measures, but it may not adapt easily to other roles.

Analytic Hierarchy Process. The AHP is a multicriterion decision making method (MCDM) that applies subjective normalized weights to decision criteria in order to rank alternatives based on decision criteria. As shown in Table 2, the AHP is a promising choice for further investigation. First, it is a logical well defined process. Specifically, the AHP uses a structured method of paired comparisons to establish weights for a hierarchy of decision criteria, or in our context, performance metrics. For example, the performance metrics (criteria) for rating a color television set might be *screen size*, *picture quality*, and *ease of use*. Then, as shown in Figure 3, *Picture quality* might be further decomposed into sub-criteria: *screen resolution* and *color saturation*. Similarly, *ease of use* might be further divided into *remote quality*, *remote compatibility*, and *on-screen programming*. Some of these criteria can be measured objectively (e.g., screen resolution) while others must be measured subjectively (e.g., remote quality).

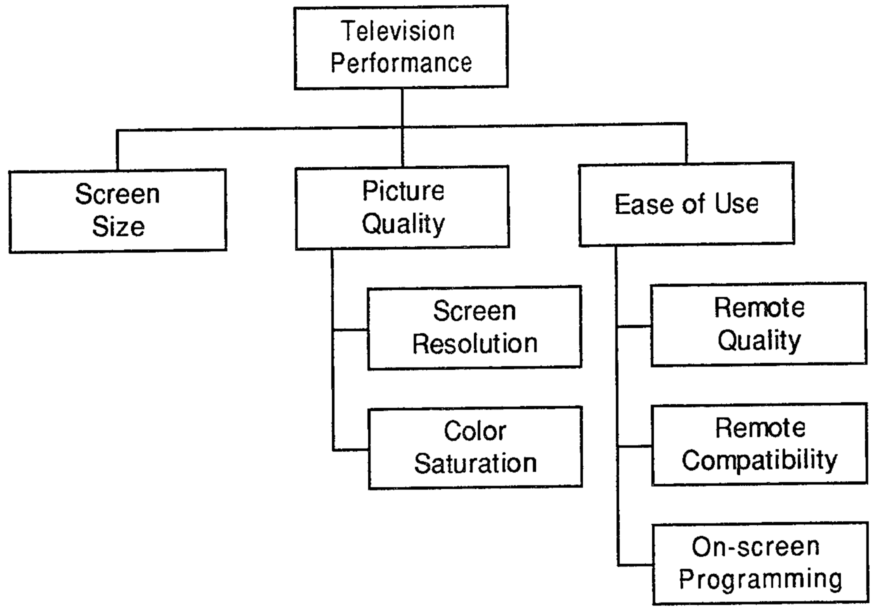



**Figure 3 Television Performance Criteria**

Objective measures can be calculated by forming a ratio of physical attributes (e.g., screen resolution / maximum screen resolution), while subjective measures are determined through paired comparisons. For example, to determine the relative weight (importance) that *remote quality, remote compatibility,* and *on-screen programming* have with respect to *ease of use,* each sub-criteria would be compared to the other two and the relative importance of each would be used to calculate a normalized weight on a zero to one scale. Similarly, each television under consideration would be objectively or subjectively compared with all other contenders, based on each of the sub-criteria. This ability to normalize any attribute is one of AHP's great strengths (Forman, 1993:19).

Second, its adaptability to a wide variety of applications is demonstrated in the literature (Saaty and Vargas, 1982:47, 103, 182, 207; Saaty 1994:427). Third, although the calculations involved with the AHP are fairly complex, commercial software is available which makes it simple to use. Fourth, it should be able to adapt to the existing set of QAF evaluation criteria. For example, Apostolou and Hassell (1993:164) used the AHP with an existing set of accounting fraud indicators in their research. While their research was not involved with organizational performance measures, it demonstrates that the AHP can adapt to an existing set of criteria (measures). Finally, although the potential for producing a USA-like score is neutral, its potential to improve the consistency and perhaps economy of the USA scoring process is good. Specifically, it may improve consistency through the process of paired comparisons which are used, like the MCP/PMT, to weight the measures for aggregation. And, unlike the MCP/PMT, paired comparisons might also be used to make subjective measurements of unit performance by comparing *actual* performance to a pre-defined definition of *ideal* performance. This paired comparison measurement ability should be a significant advantage for the USA application. However, the time required to perform the paired comparisons is probably AHP's greatest weakness (Wedley, 1993:151; Forman, 1993:25; Saaty, 1982).

As demonstrated in the preceding example, the AHP provides a well-defined, hierarchical approach to weighting higher-level criteria and lower level measures. The process for combining the resulting weights and alternative scores is also well defined. After considering all of the methods, the AHP is the best choice for further investigation.

With this decision made, we now take a closer look at both the AHP and the unit self-assessment process. Specifically, the following sections present a more thorough introduction to the AHP and the theory behind it, and an overview of the QAF-USA process.

## The Analytic Hierarchy Process

The Analytic Hierarchy Process embodies a new theory of measurement which uses paired comparisons to derive ratio scale priorities (weights or scores) for a set of decision alternatives which are compared with respect to common criteria or attributes (Vargas, 1990:2; Saaty, 1994:426). Thomas Saaty, of the Wharton School of Business, developed the AHP in the early 1970s, making it a relative newcomer to the field of decision science. Despite its youth, the AHP has attracted broad interest and found many applications in prioritization, planning, resource allocation, and prediction (Saaty and Vargas, 1982:47, 103, 182, 207). Saaty (1994:445) stresses that the purpose of the AHP is to help the individual (or group of) decision maker(s) organize their thoughts and judgments to improve the quality of a decision. It does so by organizing decision criteria into a hierarchy which is numerically weighted using a process of paired comparisons. This weighted hierarchy is then is used to compare the decision alternatives, again using a process of paired comparisons. Saaty (1986:841-842) defines this process using three generic steps:

1. Decomposition (creating the hierarchy).
2. Comparative judgments (evaluating the hierarchy).
3. Synthesis of priorities (calculating scores for the alternatives).

Decomposition is defined as the process of creating the structure of a decision problem by breaking it into logically separable elements which are organized into a hierarchy such as the one shown in Figure 4 for selecting the best ball for playing tennis. At the top of the tree a single element represents the overall *goal* of the decision, at the bottom several elements are used to represent the *alternatives* under consideration. In the middle, between the goal and the alternatives, the elements represent *criteria* that are used to evaluate either *sub-criteria* or alternatives. The many lines radiating from the alternatives connote the paired comparisons that are made with respect to every lower-most element or *leaf* of the hierarchy. Because the hierarchy structures the decision criteria, it is important that it adequately represent the problem to the satisfaction of the decision maker(s).

Once the structure has been defined, comparative judgments are made to establish the priority of criteria, sub-criteria, and alternatives. The judgments are made from the top of the hierarchy down, by comparing lower level elements, in a pairwise fashion, with respect to the goal, criteria, or sub-criteria in the next higher level. For example, the first set of paired comparisons for the example hierarchy would compare the importance (priority) of *size, bounce,* and *color* with respect to the goal of choosing a good ball for playing tennis. A fundamental scale (typically ranging from 1 to 9) is used to record pairwise judgments at every level of the hierarchy. Then a ratio scale of *local priorities* (weights) is derived from these sets (matrices) of judgments. If done manually, this process involves many iterative matrix manipulations; fortunately, the calculations have been automated in commercially available software.

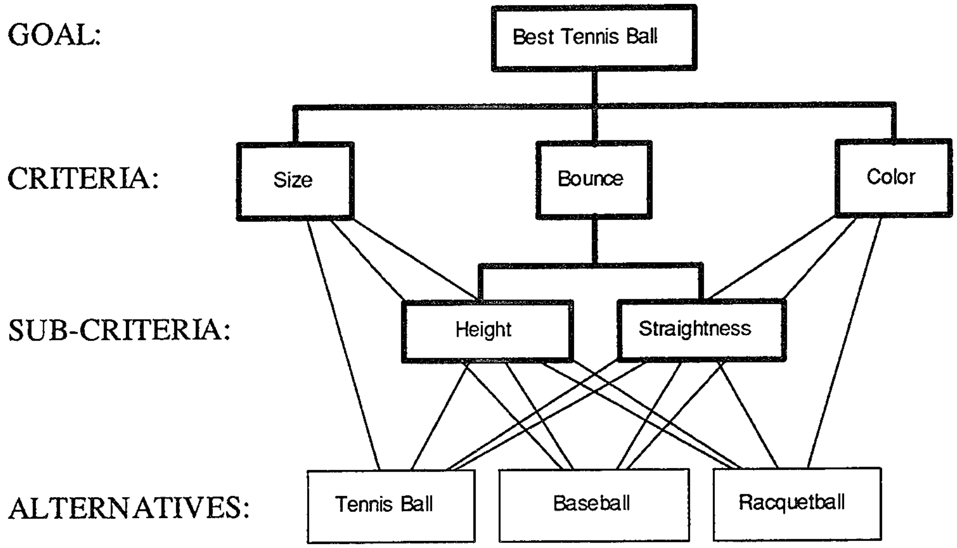



**Figure 4 Example Hierarchy**

Finally, an overall weight (priority) is calculated for each alternative by a *synthesis* of the local priorities established for each element of the hierarchy. This synthesis follows the *hierarchic composition principle* (Saaty, 1986:846) which defines the concept that dependencies flow vertically in the hierarchy, in an analogous manner to the flow of command authority down an organizational hierarchy. Thus, the final or *global* weights of the alternatives at the bottom are determined by adding the appropriate contribution of the local weights for all the elements above. When using the *distributive mode* of synthesis, the appropriate contribution is calculated by multiplying the local weight by the weight of the element one level higher. Proceeding from the top of the hierarchy downward, a portion of the overall goal weight (1.0) is distributed to each of the alternatives based on the weight of the criteria and the evaluation of the alternatives with respect to the criteria. When this process is complete, each alternative has a weight which represents its overall priority with respect to all the other alternatives.

Now, with this basic understanding of how the process works, an overview of the four axioms which underlie the theory of the AHP is appropriate. Saaty (1986:844-847) presents the axioms in a mathematically rigorous manner, while Vargas (1990:2-3) provides a simpler layman's explanation. The following explanation will borrow from each as needed to describe:

1. Reciprocal comparisons.
2. Homogenous elements.
3. Independent elements.
4. Expectations of decision maker(s).

Axiom 1 (reciprocal). This axiom requires that a decision maker be able to make comparisons between two elements (A and B) with respect to a single criteria (C) and express the strength of his or her preference. Let $P_C(A,B)$ represent the strength of preference of A over B with respect to C. This axiom also requires that the preferences $P_C(A,B)$ and $P_C(B,A)$ satisfy the reciprocal condition $P_C(A,B) = 1/P_C(B,A)$. These simple reciprocal paired comparisons form the basis for the AHP.

Axiom 2 (homogeneity). This axiom requires that elements being compared be relatively similar with respect to the criterion of comparison. For example, if *mass* were the criterion, it would be very difficult to accurately assess the relative weights of a pebble and a boulder. However, if *density* were the criterion, then the pebble and boulder may be nearly identical and thus easily compared. The AHP typically uses a fundamental scale which spans one order of magnitude (typically 1 to 9) to represent the relative preferences of two elements. If the elements span more than an order of magnitude, they may be clustered into more homogeneous clusters which share an element at their boundaries to maintain a continuity of comparison. (Saaty, 1993:7)

Axiom 3 (independence). This axiom requires that criteria be independent of the properties of the alternatives being compared. Independent means that the weight derived for a criterion (e.g., the importance of mass) with respect to a higher criterion or goal (e.g., selecting stones for a sling shot) is not affected by the alternatives (e.g., pebbles and boulders). If this axiom is violated, a *supermatrix* approach can be used to account for dependence (feedback).

Axiom 4 (expectations). This axiom requires that the decision hierarchy be complete enough to meet the expectations of the decision maker(s). In other words, a hierarchy must include all the criteria and alternatives that are deemed necessary to adequately address the decision problem. Saaty (1986:847) emphasizes that this axiom does *not* assume the decision maker is rational. No one is perfectly rational, and the AHP does not demand such perfection.

## Quality Air Force Unit Self-assessment (QAF-USA)

The QAF-USA method adopts the Malcolm Baldrige National Quality Award criteria without modification. So, like the Baldrige Award assessment, it uses a three-layer framework of evaluation criteria consisting of a set of 7 measurement *categories* which are further divided into 28 *items* and 92 *areas*. Each category is weighted by an allocation of some portion of 1000 points. Similarly, each item within a category is weighted by an allocation of the category's point value. Table 3 shows the seven categories and their point allocations. Table 3 also shows the number of items within each category. For example, breaking out category 5.0 in Table 4 shows the five items and detailed point allocations for the items within the *Management of Process Quality* category.

TABLE 3

MALCOLM BALDRIGE NATIONAL QUALITY AWARD CATEGORIES

| Category | Number of Items within the Category | Point Allocation (Weight) |
|---|---|---|
| 1.0 Leadership | 3 | 95 |
| 2.0 Information and Analysis | 3 | 75 |
| 3.0 Strategic Quality Planning | 2 | 60 |
| 4.0 Human Resource Development and Management | 5 | 150 |
| **5.0 Management of Process Quality** | **5** | **140** |
| 6.0 Quality and Operational Results | 4 | 180 |
| 7.0 Customer Focus and Satisfaction | 6 | 300 |
| TOTAL | **28** | **1000** |

TABLE 4

CATEGORY **5.0**, MANAGEMENT OF PROCESS QUALITY

| Items | Point Allocation (Weight) |
|---|---|
| 5.1 Design and Introduction of Quality Products and Services | 40 |
| 5.2 Process Management: Product and Service Production and Delivery Processes | 35 |
| 5.3 Process Management: Business Processes and Support Services | 30 |
| 5.4 Supplier Quality | 20 |
| 5.5 Quality Assessment | 15 |
| TOTAL | **140** |

The QAF-USA scoring process can be divided into five basic steps:

1. Read the unit self-assessment report.
2. Identify strengths and areas for improvement.
3. Individually score each of the 28 items using the Baldrige-based scoring guidelines.
4. Form a consensus score among a team of three to five evaluators.
5. Calculate the overall point score based on the team consensus score.

First, an evaluator must read the *unit self-assessment report.* A USA report is prepared by the unit to address each of the categories, items, and areas which will be used to assess the unit's performance towards implementing TQM practices. Specifically, the report provides the evaluator with descriptions of the unit's key business factors, such as its mission, history, products, services, customers, suppliers, competition, and other pertinent information. It also provides the evaluator with a detailed description of how the unit addresses each of the items and areas within the QAF framework.

Second, the evaluator must identify *strengths and areas for improvement* based on the information presented in the USA report. Typically, each comment is tied to a particular item and area to provide detailed feedback to the unit after the scoring process is completed.

Third, the evaluator must assign a *score* for each item within the framework. This score is based on a scale ranging from 0 to 100 percent, and is used as a multiplier for determining the number of points awarded to the unit for each item. For example, if Item 5.2 received a score of 20 percent, then seven points (0.20 x 35) would be awarded to the unit. Each item within the QAF framework can be one of two types: *approach and* *deployment* or *results*. *Approach-and-deployment*-type items are intended to gauge in what manner and how broadly the unit has established quality control processes within the organization; *results*-type items focus on how well the existing processes are working based on the quality of their resulting products or services. Thus, there are two scoring guidelines tailored to these two types. This study used only Item 5.2, an approach and deployment item, to scope the size of this investigation; therefore, only the approach and deployment scoring guidelines are shown in Table 5.

Fourth, if the range of scores given to any item vary by more than 20 percentage points across a team of evaluators, then the evaluators must meet to come to a *team consensus* on the item(s). For example, if three evaluators gave Item 5.2 (Process Management: Product and Service Production and Delivery Processes) scores of 20, 30, and 50 percent, then they would have to meet to reduce the variance within the 20 percentage point range. In this instance, the third evaluator may be persuaded to lower his or her score from 50 to 40 percent. Once a 20 percentage point range is agreed to, the individual scores are averaged to calculate the team consensus score.

Finally, the unit's overall score is calculated by multiplying each item's allocated point value by the consensus percentage point scores and then summing up the awarded points for all 28 items.

## TABLE 5

## APPROACH AND DEPLOYMENT SCORING GUIDELINES

| Observed Performance based on USA Report | Score (%) |
|---|---|
| • No system evident, anecdotal information. | 0 |
| • Beginning of a systematic approach to address the primary purposes of the item.<br>• Significant gaps still exist in deployment.<br>• Early stages of transition from reacting to preventing problems. | 10-30 |
| • Sound, systematic approach responsive to the primary purposes of the item.<br>• Fact-based improvement process in place.<br>• No major gaps in deployment, though some areas may be in early stages.<br>• More emphasis placed on problem prevention than reaction to problems. | 40-60 |
| • Sound, systematic approach responsive to the overall purposes of the item.<br>• Fact-based improvement process is a key management tool; evidence of refinement as a result of improvement cycles and analysis.<br>• Well deployed with no significant gaps, although refinement, deployment, and integration may vary among work units. | 70-90 |
| • Sound, systematic approach fully responsive to all requirements of the item.<br>• Approach is fully deployed without weaknesses or gaps in any areas.<br>• Very strong refinement and integration -- backed by excellent analysis. | 100 |

Unfortunately, as mentioned in Chapter I, ASC has discovered that the QAF-USA scoring process is not particularly consistent or economical. These flaws also weaken the QAF-USA measure when compared to an ideal metric. Ideally, the Air Force Quality Center (1993:V-3) desires metrics which are:

- Meaningful to the customer.
- Simple, understandable, logical, and repeatable.
- Able to show a trend.

- Clearly defined.
- Based on data that is economical to collect.
- Timely.
- Able to drive appropriate action.
- Able to give a snapshot of how organizational goals and objectives are being met through process and tasks.

After considering this list of ideal characteristics, it is evident that the QAF-USA method comes up short in several areas. First, it is 1) relatively unrepeatable and 2) unable to show an meaningful trend over time, both due to its inconsistency. Second, it has not proven to be very economical. And, finally, its ability to give a snapshot of performance is questionable, again due to its inconsistency.

Thus, the objective of this research is to determine whether the AHP can be used to create a new USA scoring process which is an improvement over the existing QAF-USA method. While not all of the shortcomings listed above are specifically addressed, the fundamental issues of consistency and economy are explored. The specific questions which were addressed by this research are outlined in the next section.

## Research Questions

This research assesses the feasibility and desirability of using the AHP to aggregate organizational performance metrics. Specifically, this study proposes an AHP-based approach for scoring unit self-assessment reports, and compares this new approach with the existing QAF-USA method in order to answer the following questions:

1. Is the proposed AHP-based USA scoring method feasible?
    a. Can the AHP-USA method generate an accurate aggregate score?
    b. Can the AHP adapt to the existing QAF-USA criteria?

2. Is the proposed AHP-based USA scoring method desirable?
   c. Is the AHP method more consistent than the QAF method?
   d. Is the AHP method more economical than the QAF method?
   e. Is the AHP method more understandable than the QAF method?
   f. Is the AHP method more usable than the QAF method?
   g. Is the AHP method more believable than the QAF method?
   h. Is the AHP method equally or more applicable to USA than the QAF method?

The methods used to answer these questions are defined in Chapter III. Then, each question is answered in Chapter IV.

## Summary

Measuring organizational performance has been shown to be necessary for effective management. This is true at every level of the organization. However, top management needs a significantly different level of detail than the shop floor maintenance manager. Aggregating lower level metrics into a comprehensive performance index would give a broad indicator of organizational performance that top management could use for the decision making necessary to assess and control complex processes. Although aggregate measures often obscure the source of variances, a more detailed investigation can be undertaken if the overall indicator shows an unfavorable trend. Several methods to produce an aggregated performance measure were reviewed, and the AHP was shown to be a promising approach that has been widely researched in its primary role as a multicriterion decision making tool (Vargas, 1990). However, its potential for aggregating organizational performance metrics has not yet been explored.

Chapter III presents an overview and detailed discussion of the methodology used in this study. Specifically, it reviews pre- and post-experiment steps that were used to reduce the time required to perform the scoring experiment. Also, the data collection instruments and data analysis methods are presented.

# III. Methodology

## Introduction

At this point, performance measurement, aggregation, and aggregation methods have been discussed. Also, the AHP was determined to be the best potential replacement for the QAF-USA scoring method. Thus, the remainder of this study now focuses on comparing a proposed AHP-USA scoring method to the current QAF-USA method.

First, this chapter describes how this study was performed and how the resulting data was analyzed. Specifically, it provides an overview of the methodology used to collect the data during the AHP-USA and QAF-USA scoring experiment, then it reviews pre- and post-experiment steps which were needed to implement the proposed AHP-USA scoring method. Finally, it presents the data analysis methods used to compare the results of the scoring experiment.

After presenting the methodology, Chapters IV and V present the data analysis and conclusions, respectively. In particular, Chapter IV uses the analysis methods presented in this chapter to answer each of the detailed research questions. However, the answers to the fundamental questions of feasibility and desirability are deferred to the summary and conclusions contained in Chapter V.

## Overview

At the core of the methodology is the scoring experiment used to generate data for comparing the AHP- and QAF-USA scoring methods. Figure 5 provides an overview of this experimental process:

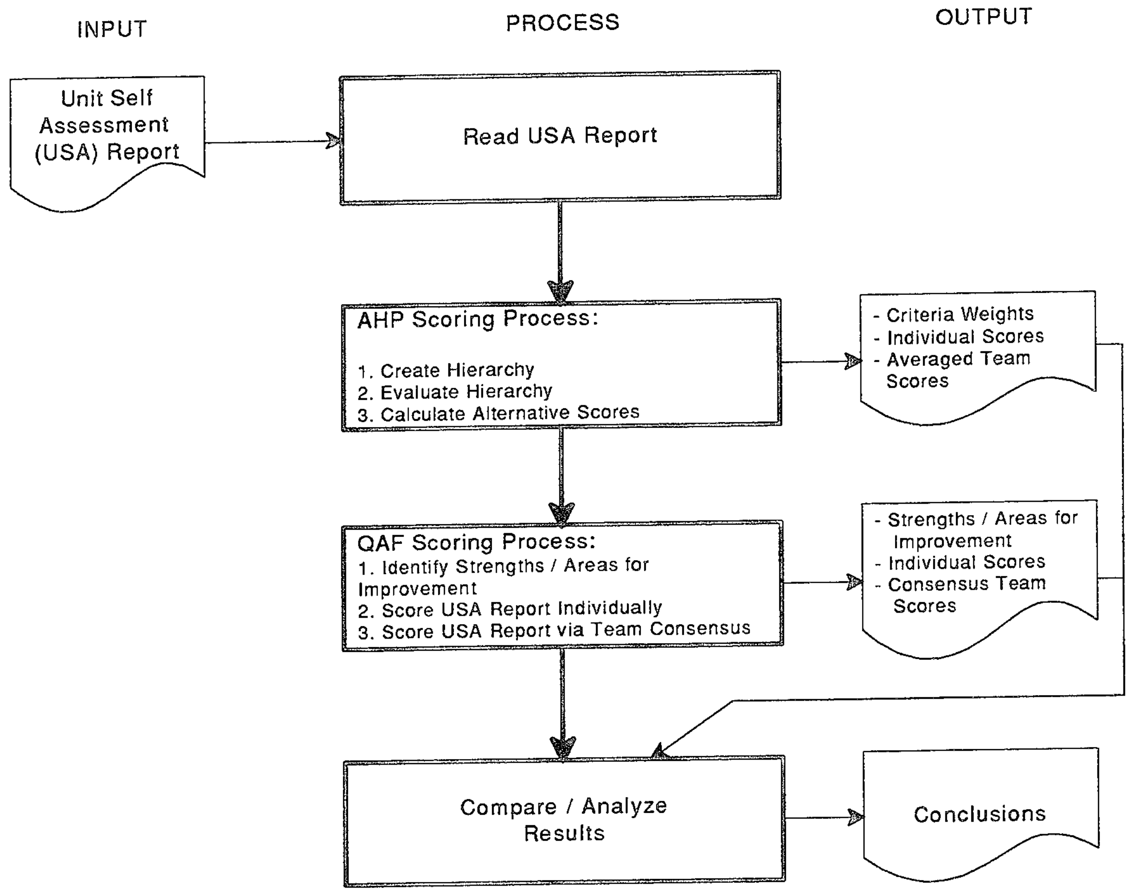



**Figure 5 Study Methodology**

It shows a simple experiment which was used to generate the data needed to compare the AHP-based scoring method to the usual QAF Baldrige-based scoring method. As shown, the input for both scoring methods consisted of an excerpt from a 1994 unit self-assessment report. A group of 11 evaluators read the report and then scored it using first the AHP-based scoring method and then the QAF scoring method. After the scoring process were complete, each evaluator was asked to answer 15 additional questions using a paired comparison approach. These questions provided feedback on the understandability, usability, believability, and applicability of the scoring methods. The results of these steps are presented in chapter IV.

The remainder of this chapter describes some pre- and post-experiment steps that helped reduce the time required to perform the experiment, and the data collection instruments and methods which were used to analyze the results.

## Pre-experiment Steps

Selecting the USA Report Excerpt. The USA report scored in the experiment consisted of an excerpt from a 1994 report written by an acquisition support unit within the Aeronautical Systems Center (ASC). Specifically, Item 5.2 (Process Management: Product and Service Production and Delivery Processes) was selected as the appropriate excerpt from the 28 possible QAF USA criteria items. This item was selected based on its complexity and relative objectivity. It provided sufficient complexity to create a relatively diverse three-level hierarchy, yet did not require an excessive amount of time to evaluate using the paired comparisons that the AHP employs. Also, based on discussions with USA experts, the items in Management of Process Quality (Category 5.0) were considered to be more objective than some of the other categories, such as Leadership (Category 1.0).

Once the item was selected, an appropriate report was selected from those on file at the ASC quality office (ASC/QI). The search was narrowed to two recent candidates based on a subjective assessment of quality which was provided by one of the ASC/QI USA experts. The final decision between the two reports was made by reviewing the reports for clarity and self-containment. In other words, the report was selected because its write-up of Item 5.2 was readable and it did not substantially cross reference other sections of the report.

Creating the AHP Hierarchy. As shown in Figure 5, this is the first step of the AHP scoring process. It involves decomposition (as described in chapter II) of the problem to create the decision hierarchy. For this experiment, the hierarchy was defined by a decomposition of Item 5.2 of the 1993 version of the QAF Criteria as shown in Figure 6. The elements of this hierarchy are defined in Table 6.

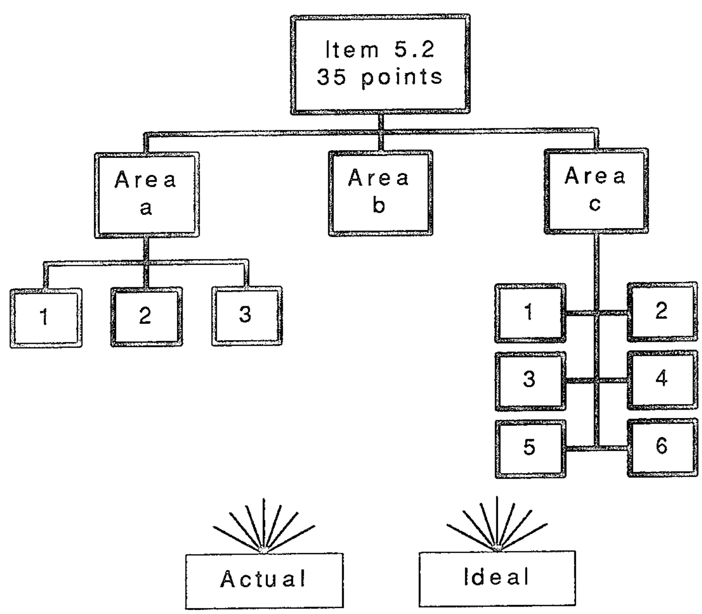



Figure 6 Item 5.2 Hierarchy

TABLE 6

DEFINITION OF ITEM 5.2 HIERARCHY ELEMENTS

| Element | Definition |
|---|---|
| Item 5.2 | Process Management: Product and Service Production and Delivery Processes |
| Area a | How quality of production and delivery services is maintained. |
| Area b | How root causes of variations in processes or outputs are determined, corrected, and verified. |
| Area c | How the process is improved for quality, cycle time, and overall operational performance. |
| Sub-area a.1 | Key processes and their requirements. |
| Sub-area a.2 | Key indicators of quality and operational performance. |
| Sub-area a.3 | How quality and operational performance are determined and maintained, including in-process and end-of-process measurements used. |
| Sub-area c.1 | Process analysis/simplification. |
| Sub-area c.2 | Benchmarking information. |
| Sub-area c.3 | Process research and testing. |
| Sub-area c.4 | Use of alternative technology. |
| Sub-area c.5 | Information from customers of the processes. |
| Sub-area c.6 | Challenge goals. |
| Actual | This alternative represents the actual performance of the unit as defined by the USA report. |
| Ideal | This alternative represents the ideal performance as defined by a perfect 100 percent score on the QAF Approach/Deployment scoring guidelines. |

## Post-experiment Steps

Calculating Individual AHP-based Scores. As shown in Figure 5, calculating scores is the last step of the AHP scoring process. It consists of two sub-steps: synthesis of priorities (as described in Chapter II) to calculate *raw scores*, and percent score calculation to determine the unit's final score. The synthesis of priorities was done using Expert Choice software. This software automated the tasks of deriving the weights for the Item 5.2 hierarchy and synthesizing the raw scores for the *ideal* and *actual* alternatives using the distributive mode of synthesis.

As previously discussed in Chapter II, the raw scores will sum to one when using the distributive synthesis mode of the AHP. Thus, a nominal value of 1.0 is allocated to the two alternatives based on the weights of the hierarchy criteria and the evaluation of the alternatives with respect to the leaves of the hierarchy. Recalling the tennis ball example, the three alternative balls would each receive a portion of the nominal goal value. For example, the tennis ball (which is perfect -- or *ideal* -- for playing tennis) might receive the lions share with a raw score of 0.5, the racquetball might be second with 0.3, and the baseball might be third with 0.2. Similarly, in the AHP-USA scoring method the *ideal* alternative (representing hypothetically perfect performance) may receive a raw score of 0.80 while the *actual* alternative (representing the unit's perceived performance) receives a raw score of 0.20.

Final percentage scores were then calculated from the raw scores by forming the ratio ACTUAL/IDEAL. Thus, the percentage score for the previous example would be calculated as follows: 0.20/0.80 = 0.25 or 25 percent. The rationale for calculating a percentage score in this manner derives from the definition of the *ideal* alternative used in the AHP-USA method. Because the *ideal* alternative represents a unit which deserves a perfect score of 100 percent (1.0) on the QAF approach/deployment scale (see Table 5), the *actual* alternative's percentage score is determined by solving Equation (2) for $X$, where $X$ represents the unit's actual percentage score on the QAF scale:

$$\frac{X}{100} = \frac{Actual}{Ideal} \tag{2}$$

Solving for $X$ simply translates the raw scores generated by the AHP-USA method (i.e., 0.20 and 0.80) into an equivalent percentage point score (i.e., 25 percent) that matches the 0-100 scale used in the QAF-USA method.

Calculating Team AHP-based Scores. Calculating team scores is very similar to calculating individual scores. The only difference being that the geometric means of the team members' individual paired comparisons were used to weight and synthesize the hierarchy. Saaty (1980:227) states that the geometric mean can be used to combine individual judgments when debate has not resulted in consensus. This study deliberately avoided using a consensus process for the AHP-USA scoring method. There were two reasons for this: 1) including consensus steps as part of the AHP-USA method would require more time to use the method, thus working against the objective of finding a more economical scoring process, and 2) it was desirable to limit the participation time required of the evaluators. Thus, it was decided to weight *team* hierarchies by using the geometric means of the team members' individual paired comparison judgments.

To demonstrate the team score calculations, consider a team of four evaluators and the previous tennis ball example. First, each evaluator completes all of the paired comparisons necessary to evaluate the complete hierarchy. This results in a set of four judgments for each comparison of criteria and alternatives. One of these sets is shown in Figure 7.

| Which criteria (size, bounce, and color) is more important with respect to the goal? | | | | | |
|---|---|---|---|---|---|
| (Goal): Best tennis ball. | | | | | |
| Eval # | | | | | |
| 1 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 2 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 3 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 4 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |

Figure 7 Team's Set of Judgments

These judgments are simply mapped into the appropriate preference value by using the full AHP 1 to 9 scale and reciprocals as appropriate. Table 7 summarizes this mapping.

TABLE 7

PREFERENCE SCALE

| Degree of Preference | Preference Value | |
|---|---|---|
| | A preferred over B (Left over Right) | B preferred over A (Right over Left) |
| EXTREME | 9 | 1/9 |
| Very Strong to Extreme | 8 | 1/8 |
| VERY STRONG | 7 | 1/7 |
| Strong to Very Strong | 6 | 1/6 |
| STRONG | 5 | 1/5 |
| Moderate to Strong | 4 | 1/4 |
| MODERATE | 3 | 1/3 |
| Equal to Moderate | 2 | 1/2 |
| EQUAL | 1 | 1 |

Another way to think of this mapping is simply to ask the question "How much is A preferred over B?" If the answer is "A is extremely preferred over B" or "A is preferred nine times more than B," then the preference value is 9. On the other hand, if the answer is "A is extremely *not* preferred to B" or "A is preferred one-ninth as much as B," then the preference value is 1/9.

Continuing with the example, the judgments shown in Figure 7 are mapped into their corresponding values and their geometric mean is calculated as shown in Table 8.

**TABLE 8**

**GEOMETRIC MEAN**

| Evaluator 1 | Evaluator 2 | Evaluator 3 | Evaluator 4 | Geometric Mean | Reciprocal of Geometric Mean |
|---|---|---|---|---|---|
| 1 | 1/2 | 1/7 | 2 | 0.615 | 1.627 |

Considering that the geometric mean falls between 1/2 and 1 in Table 7, it is easy to see that the team prefers the right-hand criteria (bounce) over the left-hand criteria (size). The degree of this *right over left* preference can be determined by taking the reciprocal of the geometric mean, giving a value of 1.627 in favor of *bounce*.

## Data Collection Instrument

A structured briefing and coordinated data collection package were used to guide the experiment and record data from the 11 evaluators. A copy of the briefing and data collection package can be found in Appendices A and B, respectively. The briefing and data collection package were used to:

1. Instruct the evaluators in the AHP scoring method.
2. Review the steps in the QAF scoring method.
3. Collect demographic data about the evaluators.
4. Collect the AHP paired comparison judgments for the Item 5.2 hierarchy.
5. Collect individual and team consensus scores from the QAF scoring process.
6. Collect elapsed times for the two methods.
7. Collect scoring process feedback using paired comparisons and open-ended comments.

Functions 1, 2, and 3 above are self-explanatory. The development of the data collection package to perform functions 4 through 7 are described below.

The worksheets used for collection of the paired comparison evaluations (function 4) were based on the format of questionnaires that can be generated using Expert Choice. Also, Saaty (1980:34) proposes using questionnaires similar to the example shown in Figure 8. Each worksheet presents the paired comparisons to the evaluator and provides the semantic scale at the bottom of each page. Computer-based data collection would have been preferable for this study because it would have provided immediate feedback to the evaluators regarding the consistency of their judgments. However, neither the hardware nor software resources were available to collect the data by computer.

The worksheets used for collection of the individual and team consensus scores for the QAF scoring process (function 5) were developed based on similar forms employed by ASC/QI. These forms can be found in Appendix B.

Elapsed times (function 6) are calculated from the start and stop times which were recorded in the spaces provided on the forms as shown in Figure 8. Times were recorded in this manner for both the AHP and QAF scoring methods. However, the time evaluators used to read the USA report was not recorded because this reading time would be equivalent for both scoring methods.

Finally, another set of paired comparison worksheets was used to record evaluators' judgments of preference between the two scoring methods with respect to 15 individual questions. The questions fell into four general categories related to attributes of the scoring methods:

1. Understandability
2. Usability
3. Believability
4. Applicability

---

PLEASE RECORD THE TIME WHEN YOU BEGIN: _________

Which **area** (a, b, and c) is more important with respect to **Item 5.2**?

**(5.2): Process Management: Product and Service Production and Delivery Processes**

| | | | | |
|---|---|---|---|---|
| **(a):** How quality of production and delivery services is maintained | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(b):** How root causes of variations in processes or outputs are determined, corrected and verified |
| **(a):** How quality of production and delivery services is maintained | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(c):** How the process is improved for quality, cycle time, and overall operational performance |
| **(b):** How root causes of variations in processes or outputs are determined, corrected and verified | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(c):** How the process is improved for quality, cycle time, and overall operational performance |

PLEASE RECORD THE TIME WHEN YOU ARE DONE: _________

Degrees of Preference: **1** = Equal **3** = Moderate **5** = Strong **7** = Very Strong **9** = Extreme (2, 4, 6, and 8 are in-between values)

---

**Figure 8 Example Paired Comparison Worksheet**

After this set of questions, the worksheet solicited open ended comments from the evaluators. Again, a worksheet can be found in Appendix B. Having reviewed the data collection instrument, it is time to return to the research questions and an overview of the analyses that will be used to answer them.

## Answering the Research Questions

Here again are the research questions as proposed in Chapter II with an added outline of the detailed measurement questions used to address the research questions.

1. Is the proposed AHP-based USA scoring method feasible?
   a. Can the AHP-USA method generate an accurate aggregate score?
      (1) Are the individual scores generated by both methods equivalent?
      (2) Are the team scores generated by both methods equivalent?
      (3) Are the team scores generated by both methods equivalent to the historical QAF-USA score given to the unit originally?
   b. Can the AHP adapt to the existing QAF-USA criteria?
      (1) Can an appropriate AHP hierarchy be constructed from the QAF criteria?
      (2) Can evaluators consistently judge the resulting hierarchy?
2. Is the proposed AHP-based USA scoring method desirable?
   a. Is the AHP method more consistent than the QAF method?
      (1) Is the range of individual scores from AHP-USA less than the range of individual scores from the QAF-USA method?
   b. Is the AHP method more economical than the QAF method?
      (1) Are the elapsed times for the AHP-USA method less than the elapsed times for the QAF-USA method?
   c. Is the AHP method more understandable than the QAF method?
      (1) Which method is easier to understand?
      (2) Which method is more intuitive?
      (3) Which method is simpler?
   d. Is the AHP method more usable than the QAF method?
      (1) For which method was it easier to determine a score for the unit?
      (2) Which method was easier to use overall in this scoring exercise?
      (3) Which method was a better tool to identify strengths and areas for improvement for the unit?
   e. Is the AHP method more believable than the QAF method?
      (1) If an "accurate score" is defined as "a score which truly represents a unit's performance relative to desired/planned performance," which method do you believe would produce the most accurate scores?
      (2) If a "consistent score" is defined as "a score which varies little between evaluators," which method do you believe would produce the most consistent scores?
      (3) Which method would you trust more to score unit self-assessments?
   f. Is the AHP method equally or more applicable to USA than the QAF method?

(1) Which method would you prefer to use for future unit self-assessments?
(2) Which method would you prefer to accurately show a trend in USA scores from year to year?
(3) Which method would you prefer if USA were to be done by novice practitioners?
(4) Which method would you prefer if USA were to be done by experienced practitioners?
(5) Which method is best suited to a variety of USA practitioners (novice to experienced)?
(6) Which method do you prefer overall?

The two fundamental questions are whether the proposed AHP-USA scoring method is 1) feasible, and 2) desirable. These basic research questions were broken into eight operational questions (1.a. through 2.f) and 22 measurement questions (1.a.(1) through 2.f.(6)). Several different analysis methods were used to answer these questions. The remainder of this chapter presents these methods.

## RMS, MAD, and MAPE

Saaty (1980:37-38) uses the root mean square deviation (RMS) and the median absolute deviation (MAD) about the median to judge the accuracy of a *measurement vector* (i.e., set of measurements) with respect to a *true vector* (i.e., a set of values which are known to be correct or *true*). In this study, the measurement vectors were sets of unit scores (both individual and team) derived from the AHP-USA scoring method. Sets of comparable scores derived using the conventional QAF-USA scoring method were used as the true vectors. Of course, using the QAF-USA scores as the reference is not ideal due to the inconsistency of the QAF-USA method. Unfortunately, no better measure of actual unit performance is available.

Equation (3) gives the definition of the RMS for two vectors (sets) of numbers denoted by $a_1, a_2,\ldots,a_n$ and $b_1, b_2,\ldots,b_n$.

$$\sqrt{\frac{1}{n}\sum_{i-1}^{n}(a_i - b_i)^2} \tag{3}$$

Equation (4) gives the definition of the MAD for two vectors (sets) of numbers also denoted by $a_1, a_2,\ldots,a_n$ and $b_1, b_2,\ldots,b_n$.

$$\text{median}\left[\left|(a_i - b_i) - \text{median}(a_i - b_i)\right|\right] \tag{4}$$

The median of a set of $n$ numbers is determined by sorting the numbers to obtain an ascending list. If $n$ is odd, then the median is the value at the exact middle of the sorted list. If $n$ is even, then the median is the average of the two numbers which straddle the exact middle.

The RMS and MAD are not used directly to evaluate the accuracy of the measurement vector. Instead, Saaty (1980:39) calculates two significance ratios by dividing the RMS and the MAD by the average value of the vector components. In his application, the components of the vectors always sum to one, thus the average value is simply $1/n$, where $n$ is the number of components. This is not the case for the sets of unit scores in this study. Therefore, the significance ratios ($S_{RMS}$ and $S_{MAD}$) were calculated by using the average value of both sets of vector components as shown in Equations (5) and (6) below.

$$S_{RMS} = \frac{RMS}{\frac{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i}{2n}} \tag{5}$$

$$S_{MAD} = \frac{MAD}{\frac{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i}{2n}} \tag{6}$$

Saaty (1980:39) states that "two vectors are nearly the same if either or both ratios are...less than 0.1." Thus, if either $S_{RMS}$ or $S_{MAD}$, or both are less than 0.1, then the sets of unit scores will be considered equivalent.

Finally, the mean absolute percentage error (MAPE) is also used to assess the relative accuracy of the AHP-USA scores when compared to the QAF-USA scores. Like the significance ratios above, the MAPE is independent of the magnitude of the vectors being compared (Nahmias, 1993:57). The formula for the MAPE, using the same definitions of *a* and *b* vectors as defined above, is shown in Equation (7).

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{a_i - b_i}{b_i}\right|\right] \times 100 \tag{7}$$

For two vectors to be equivalent, the MAPE should be less than 10.

## Consistency Ratios

The AHP provides a way to measure whether the paired comparison judgments were made with logical consistency. A measure of consistency is generated by calculating an AHP *consistency ratio* as described by Saaty (1982:84). The higher the consistency ratio, the more inconsistent the paired comparison judgments. For example, recall the top level criteria for selecting the best ball for playing tennis: size, bounce and color. Assume that two different evaluators have completed the paired comparison worksheets as shown in Figure 9 and Figure 10.

| Which criteria (size, bounce, and color) is more important with respect to the goal? | | | | |
|---|---|---|---|---|
| (Goal): Best tennis ball. | | | | |
| Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Color |
| Bounce | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Color |

Figure 9 Evaluator A's Judgments

| Which criteria (size, bounce, and color) is more important with respect to the goal? | | | | |
|---|---|---|---|---|
| (Goal): Best tennis ball. | | | | |
| Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Color |
| Bounce | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Color |

Figure 10 Evaluator B's Judgments

Both evaluators are inconsistent in their judgments, but to different degrees. Specifically, both have judged that *size* and *bounce* are of equal importance with respect to choosing a ball for playing tennis; however, both evaluators subsequently judge that *size* is of lesser importance than *bounce* when they are compared with *color*. Thus, both evaluators are inconsistent because if size and bounce are truly equal, then they should be equal when compared with color. Furthermore, the degree to which they are inconsistent is reflected in the consistency ratios as shown in Table 9.

**TABLE 9**

**CONSISTENCY RATIOS**

| Evaluator | Consistency Ratio |
|---|---|
| A | 0.007 |
| B | 0.130 |

A consistency ratio greater than 0.1 indicates that the paired comparison judgments may be too random (Saaty, 1982:83). So, evaluator B above is slightly over Saaty's threshold for good consistency. Usually, if this situation occurs, the inconsistency is brought to the attention of the evaluator and the inconsistent comparisons are revised in an effort to reduce the consistency ratio. However, this was not done in this study. Instead, the consistency ratios were used as an indicator of whether the evaluators were able to consistently compare the criteria that were derived from Item 5.2 of the QAF-USA evaluation criteria.

## Histograms

For all of the feedback paired comparisons, the frequency of the evaluators' responses are plotted as bars on the y-axis versus the double ended 1 to 9 scale along the x-axis. Also, the geometric mean of the responses is indicated with a text box and arrow to the x-axis. These histograms provide a *picture* of the evaluators responses. Qualitative conclusions are then drawn from an analysis of these graphs. For example, the 11 judgments shown in Figure 11 can be processed using a spreadsheet and then plotted as shown in Figure 12.

| Which criteria (size, bounce, and color) is more important with respect to the goal? | | | | | |
|---|---|---|---|---|---|
| (Goal): Best tennis ball. | | | | | |
| Eval # | | | | | |
| 1 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 2 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 3 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 4 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 5 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 6 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 7 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 8 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 9 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 10 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| 11 | Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |

Figure 11 Set of Paired Comparison Judgments

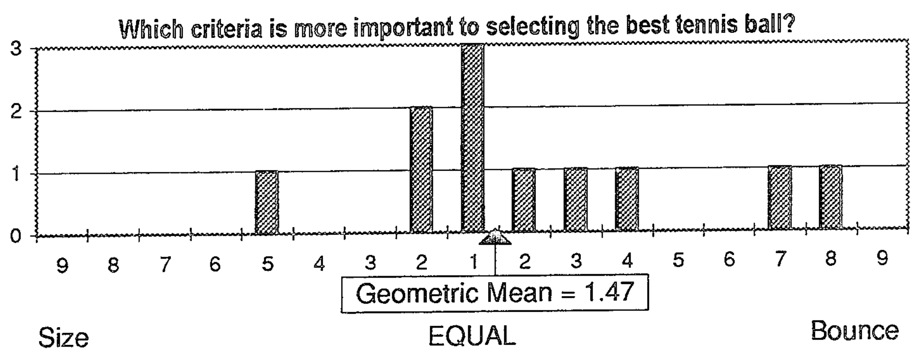



Figure 12 Histogram with Geometric Mean

The bars on the graph show how the evaluators judged the criteria and the triangle just to the right of "1" on the x-axis shows the location of the geometric mean calculated to be 1.47 towards *bounce*. Thus, this chart shows that the team of evaluators judged *bounce* to be slightly more important than *size*; however, the variance in opinion is fairly large. Due to the variance and the geometric mean being fairly close to 1.0, it is difficult to make a strong statement of preference in this example. Questions 2.c through 2.f are answered in this manner.

Chapter IV describes the data and analyses of this study. Specifically, it further describes the unit self-assessment report, the demographics of the participating evaluators, and summarizes the results of the data collection. Then, it presents the analyses performed to answer each research question.

# IV. Data Description, Analysis, and Findings

This chapter presents the *raw* input, output, and process-related data which is pertinent to the analysis and subsequent findings. The data description section strives to present this material in a comprehensive, yet compact, manner. This chapter also presents the *processed* results of the analyses described in Chapter III.

## Data Description

Unit Self-assessment (USA) Report. As described in Chapter III, this report was the primary input to the USA scoring processes. It consisted of an excerpt from a complete 1994 USA report that was prepared by one of the acquisition support units within Aeronautical Systems Center (ASC). For purposes of non-attribution, the name of the unit is not provided here; however, the USA report states that the unit is comprised of approximately 90 government employees (40 percent military) with an annual budget of about $30 million. Within ASC's organizational structure, the unit resides at a level equal to the major system program offices (SPOs), such as the F-22, B-2, C-17, and F-16 SPOs. Internally, the unit is comprised of its commander, four integrated product teams (IPTs), a requirements division which interfaces with customers, and three functional divisions which support the IPTs.

The eight-page excerpt used in this study provided the evaluators with the unit's Key Business Factors (KBFs), its write-up of Item 5.2 (Process Management: Product and

Service Production and Delivery Processes), and a list of acronyms. The KBF section consisted of four and one-half pages which were devoted to the following topics:

1. Business description
2. The background of the unit
3. Key requirements for products and services
4. Customer base
5. Markets
6. Competitive environment
7. Major equipment, facilities and technologies
8. Suppliers
9. Relationships with customers and suppliers
10. Regulatory environment
11. The future of the unit
12. Link to parent organization

Other than a short list of acronyms, the remainder of the document addressed the three areas within Item 5.2. Thus, about three pages were devoted to explaining how the unit 1) maintains the quality of its production and delivery services, 2) determines, corrects, and verifies the root causes of variation in processes or outputs, and 3) improves its processes for quality, cycle time, and overall operational performance. This narrative information is typical of a USA report, and is the primary source of information from which evaluators must determine a score for the unit.

Supplemental Reference to the USA Report. In addition to the USA report, another document containing selected excerpts of the full USA report was made available to the evaluators. This was done because many of the QAF/Baldrige criteria are interrelated. For example, Area 5.2.a. refers to the "product and service design requirements" which are to be identified in Item 5.1 (AFQI, 1993:21). Similarly, Brown (1993:56-67) identifies other interrelationships which evaluators may wish to consider when scoring a USA report. Therefore, the supplemental reference contained the following sections from the full USA report:

1. Section 2.0 - Information and Analysis
2. Section 2.1 - Scope and Management of Quality and Performance Data and Information
3. Section 2.2 - Competitive Comparisons and Benchmarking
4. Section 5.0 - Management of Process Quality
5. Section 5.1 - Design and Introduction of Quality Products and Services

Based on casual observations during data collection, none of the evaluators used this supplemental reference to aid their judgments, therefore, further description of these items is not provided here.

Evaluator Demographics. Eleven people participated as evaluators in this research. They evaluated (scored) the USA report described above using both the AHP-based and the QAF scoring methods. They also made paired comparison judgments to provide feedback on both scoring processes, and provided unstructured narrative comments as well.

The participants were selected based primarily on their training and experience with the QAF-USA scoring method; however, four participants had no formal USA training, and three of these also had no experience with the QAF-USA scoring process. These novice participants who were unfamiliar with either scoring process may have provided a less biased and, granted, a less focused, perspective. Overall, however, the panel of evaluators had significant training and experience as summarized in Table 10.

## TABLE 10

## TEAM AND EVALUATOR DEMOGRAPHICS

| Teams: | A | | | | B | | | | C | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Evaluators: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| **USA Training** | | | | | | | | | | | |
| < 3 months | | | | | | | | | | X | |
| 3-6 months | | | | | | | | | | | |
| 6-12 months | | X | X | | | | | | | | |
| > 12 months | X | X | | | X | X | | | X | | |
| **USA Experience** | | | | | | | | | | | |
| Data collection | X | X | X | X | X | X | | | X | | |
| Scoring | X | X | | | X | X | | | | | |
| Consultant | X | X | | | X | X | | | | | |
| Examiner | X | | | | X | | | | | | |
| Other | X | | | X | X | | | | | | |
| **Training or Experience** | X | X | X | X | X | X | | | X | X | |

For training, Table 10 simply shows whether the individual evaluator was formally trained on the USA process, and when training had been conducted. To gather data on their experience, the participants were asked to mark all of the following statements which applied:

- Participated in data collection/preparation of self-assessment reports for your unit.
- Scored USA reports.
- Visited unit(s) as a USA consultant (for feedback to the unit).
- Visited unit(s) as a USA examiner (for award application site visit).
- Other: ___________________________________

For those evaluators that marked *other* and filled in the blank, Table 11 summarizes their comments about their experience.

TABLE 11

EXPERIENCE COMMENTS

| Evaluator | Comments |
|---|---|
| 1 | Involved with Quality Dayton Award process.<br>Train [others] on USA and QAF criteria. |
| 4 | Compiled and presented [USA] information to higher authority. |
| 5 | Provide training on USA.<br>Act as a consultant for center level [ASC-level] (leaders) team. |

The reader may perceive that Table 10 seems to be intentionally ordered by experience within teams, and perhaps even across teams. The perception is correct within teams, but not across teams. Specifically, during the scoring experiment, individuals were arbitrarily assigned to one of the three teams. Thus, it is only by chance that team A is the most trained and experienced, with team B in second, and team C in third. However, the evaluators were assigned numbers *after* the response packages had been sorted by experience *within* the teams; thus, the intentional pattern arises.

Results from Original USA Scoring. Only one value was needed from the original (historical) scoring of the USA report. This value was the team consensus percentage score for Item 5.2, which has a value of 52.5 percent. This was the only number needed because this study was designed to use the criteria from a single QAF item -- namely, Item 5.2. Also, only the team score (versus any individual score) was needed because it makes little sense to try to match individual scores from different evaluators. Thus, only the original team score of 52.5 percent for Item 5.2 was compared to the team scores obtained during this study.

Results from AHP-USA Scoring. This section presents a portion of the data generated by the AHP-USA scoring method. The remainder of the AHP-USA data is presented at the appropriate places within the analysis section of this chapter.

Table 12 presents the individual paired comparison judgments from all 11 evaluators for each of the Item 5.2 hierarchy criteria comparisons. The numbers in the table show the preference of the left element over the right element in accordance with the AHP preference scale (Table 7). For example, looking at the "a-b" comparison, evaluator number one had a *strong* preference for area "a" over area "b," thus a 5 is used to record this judgment. However, evaluator number four had the exact opposite preference, strongly preferring area "b" over area "a," thus the reciprocal value of 1/5 or 0.200 is used to record this judgment. The table also shows the composition of teams A, B, and C.

Table 13, in a manner similar to Table 12, presents the paired comparison judgments for the alternatives of the hierarchy with respect to the leaves of the hierarchy. Unlike Table 12, the left column of the table simply shows the single *leaf* that is being used to make the *actual-ideal* comparison. For example, with respect to leaf a.1, evaluator number one expressed a *very strong* preference for the *ideal* performance over the unit's *actual* performance. Thus, this judgment is recorded using 1/7 or 0.143.

Table 14 presents the team *judgments* for each of the three teams. *Judgments* is somewhat of a misnomer because these values are calculated by taking the geometric mean of the individual team members' individual judgments. The reciprocal of the geometric means are also presented because Expert Choice requires judgments be entered

TABLE 12

PAIRED COMPARISON JUDGMENTS FOR CRITERIA

| Eval # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Team | A | | | | B | | | | C | | |
| a-b | 5.000 | 3.000 | 5.000 | 0.200 | 5.000 | 7.000 | 0.500 | 5.000 | 1.000 | 5.000 | 5.000 |
| a-c | 5.000 | 2.000 | 0.200 | 0.200 | 1.000 | 7.000 | 0.500 | 0.200 | 1.000 | 1.000 | 3.000 |
| b-c | 3.000 | 1.000 | 0.143 | 5.000 | 0.200 | 0.200 | 1.000 | 0.200 | 1.000 | 1.000 | 4.000 |
| a.1-a.2 | 3.000 | 5.000 | 0.333 | 0.200 | 5.000 | 0.200 | 3.000 | 1.000 | 5.000 | 0.333 | 0.333 |
| a.1-a.3 | 3.000 | 3.000 | 0.333 | 0.143 | 0.333 | 0.200 | 0.333 | 1.000 | 3.000 | 0.333 | 0.333 |
| a.2-a.3 | 2.000 | 2.000 | 3.000 | 0.200 | 1.000 | 5.000 | 0.500 | 1.000 | 3.000 | 0.333 | 1.000 |
| c.1-c.2 | 5.000 | 5.000 | 5.000 | 1.000 | 3.000 | 5.000 | 5.000 | 5.000 | 3.000 | 1.000 | 3.000 |
| c.1-c.3 | 5.000 | 3.000 | 5.000 | 0.111 | 0.500 | 5.000 | 1.000 | 0.143 | 1.000 | 1.000 | 0.200 |
| c.1-c.4 | 2.000 | 3.000 | 7.000 | 3.000 | 1.000 | 9.000 | 3.000 | 0.143 | 3.000 | 0.333 | 0.200 |
| c.1-c.5 | 0.143 | 1.000 | 0.333 | 0.111 | 0.500 | 9.000 | 0.333 | 0.143 | 1.000 | 0.200 | 0.333 |
| c.1-c.6 | 1.000 | 2.000 | 6.000 | 7.000 | 3.000 | 9.000 | 4.000 | 0.200 | 1.000 | 0.333 | 3.000 |
| c.2-c.3 | 2.000 | 1.000 | 2.000 | 0.143 | 0.333 | 1.000 | 0.333 | 0.143 | 1.000 | 1.000 | 0.333 |
| c.2-c.4 | 0.167 | 1.000 | 4.000 | 0.500 | 1.000 | 1.000 | 1.000 | 0.143 | 3.000 | 0.333 | 0.250 |
| c.2-c.5 | 0.111 | 0.333 | 0.333 | 0.111 | 1.000 | 0.200 | 0.500 | 0.143 | 0.333 | 0.200 | 0.500 |
| c.2-c.6 | 1.000 | 1.000 | 5.000 | 1.000 | 5.000 | 5.000 | 3.000 | 1.000 | 3.000 | 0.333 | 1.000 |
| c.3-c.4 | 0.200 | 1.000 | 3.000 | 7.000 | 3.000 | 3.000 | 1.000 | 7.000 | 3.000 | 0.333 | 3.000 |
| c.3-c.5 | 0.111 | 0.333 | 0.200 | 5.000 | 5.000 | 0.200 | 1.000 | 1.000 | 0.333 | 0.200 | 3.000 |
| c.3-c.6 | 0.200 | 1.000 | 2.000 | 5.000 | 5.000 | 5.000 | 3.000 | 7.000 | 1.000 | 1.000 | 5.000 |
| c.4-c.5 | 0.500 | 0.333 | 0.200 | 0.333 | 0.500 | 0.200 | 0.500 | 1.000 | 0.333 | 0.333 | 5.000 |
| c.4-c.6 | 2.000 | 1.000 | 1.000 | 1.000 | 0.500 | 1.000 | 2.000 | 7.000 | 0.333 | 1.000 | 5.000 |
| c.5-c.6 | 7.000 | 3.000 | 8.000 | 5.000 | 3.000 | 5.000 | 3.000 | 7.000 | 5.000 | 3.000 | 3.000 |

TABLE 13

PAIRED COMPARISON JUDGMENTS FOR ALTERNATIVES

| Eval # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Team | A | | | | B | | | | C | | |
| a.1 | 0.143 | 0.200 | 0.167 | 0.333 | 0.200 | 0.333 | 0.333 | 1.000 | 0.200 | 0.200 | 0.200 |
| a.2 | 0.125 | 0.200 | 0.143 | 1.000 | 0.200 | 0.200 | 0.250 | 1.000 | 0.200 | 0.200 | 0.143 |
| a.3 | 0.111 | 0.143 | 0.143 | 1.000 | 0.125 | 0.333 | 0.250 | 1.000 | 0.200 | 0.333 | 0.333 |
| b | 0.111 | 0.111 | 0.125 | 0.333 | 0.111 | 0.500 | 0.200 | 1.000 | 0.143 | 0.333 | 0.143 |
| c.1 | 0.125 | 0.200 | 0.143 | 0.333 | 0.125 | 1.000 | 0.167 | 0.333 | 0.143 | 0.200 | 0.167 |
| c.2 | 0.111 | 0.125 | 0.167 | 0.333 | 0.143 | 0.200 | 0.250 | 1.000 | 0.200 | 0.500 | 0.143 |
| c.3 | 0.111 | 0.111 | 0.143 | 1.000 | 0.111 | 0.333 | 0.500 | 0.200 | 0.200 | 0.200 | 0.333 |
| c.4 | 0.143 | 0.200 | 0.143 | 1.000 | 0.333 | 0.200 | 1.000 | 0.333 | 0.200 | 0.200 | 0.143 |
| c.5 | 0.167 | 0.200 | 0.167 | 1.000 | 0.200 | 0.200 | 0.143 | 0.143 | 0.143 | 0.200 | 0.333 |
| c.6 | 0.143 | 0.125 | 0.143 | 0.200 | 0.143 | 0.200 | 1.000 | 0.333 | 0.143 | 0.333 | 0.167 |

using numbers which are greater than or equal to one, with the option of reversing the order of the preference. For example, in order to correctly enter team A's geometric mean for the "a-c" comparison, 1.257 is entered into Expert Choice and the direction of the preference is reversed (i.e., "c" over "a" rather than "a" over "c"). This artifact of the Expert Choice software is the only reason for presenting the reciprocals of the geometric means.

Table 15 presents the calculated local priority weights for each of the criteria in the Item 5.2 hierarchy. These weights were calculated by the Expert Choice software based on the paired comparison judgments contained in Table 12. For example, looking back at Table 12, it is easy to see that evaluator one preferred "a" over "b," "b" over "c," and "a" over "c" when comparing the level 1 criteria. Therefore we would expect that "a" should receive the most weight, "b" should receive the second largest allocation, and "c" (which was dominated by both "a" and "b") should receive the smallest allocation of weight. From Table 15, evaluator number one's calculated weights for these criteria are 0.701, 0.202, and 0.097, which qualitatively matches our expectations. Also, as shown by the summary rows, the local priority weights properly sum to 1.0 within the limits of round-off error.

TABLE 14

TEAM JUDGMENTS CALCULATED USING THE GEOMETRIC MEAN

| Criteria | Team A | | Team B | | Team C | |
|---|---|---|---|---|---|---|
| | Geometric Mean (GM) | Reciprocal (1/GM) | Geometric Mean (GM) | Reciprocal (1/GM) | Geometric Mean (GM) | Reciprocal (1/GM) |
| a-b | 1.968 | 0.508 | 3.058 | 0.327 | 2.924 | 0.342 |
| a-c | 0.795 | 1.257 | 0.915 | 1.093 | 1.442 | 0.693 |
| b-c | 1.210 | 0.827 | 0.299 | 3.344 | 1.587 | 0.630 |
| a.1-a.2 | 1.000 | 1.000 | 1.316 | 0.760 | 0.822 | 1.216 |
| a.1-a.3 | 0.809 | 1.236 | 0.386 | 2.590 | 0.693 | 1.442 |
| a.2-a.3 | 1.245 | 0.803 | 1.257 | 0.795 | 1.000 | 1.000 |
| c.1-c.2 | 3.344 | 0.299 | 4.401 | 0.227 | 2.080 | 0.481 |
| c.1-c.3 | 1.699 | 0.589 | 0.773 | 1.294 | 0.585 | 1.710 |
| c.1-c.4 | 3.350 | 0.298 | 1.401 | 0.714 | 0.585 | 1.710 |
| c.1-c.5 | 0.270 | 3.708 | 0.680 | 1.470 | 0.405 | 2.466 |
| c.1-c.6 | 3.027 | 0.330 | 2.156 | 0.464 | 1.000 | 1.000 |
| c.2-c.3 | 0.869 | 1.150 | 0.355 | 2.817 | 0.693 | 1.442 |
| c.2-c.4 | 0.760 | 1.316 | 0.615 | 1.627 | 0.630 | 1.587 |
| c.2-c.5 | 0.192 | 5.196 | 0.346 | 2.893 | 0.322 | 3.107 |
| c.2-c.6 | 1.495 | 0.669 | 2.943 | 0.340 | 1.000 | 1.000 |
| c.3-c.4 | 1.432 | 0.699 | 2.817 | 0.355 | 1.442 | 0.693 |
| c.3-c.5 | 0.439 | 2.280 | 1.000 | 1.000 | 0.585 | 1.710 |
| c.3-c.6 | 1.189 | 0.841 | 4.787 | 0.209 | 1.710 | 0.585 |
| c.4-c.5 | 0.325 | 3.080 | 0.473 | 2.115 | 0.822 | 1.216 |
| c.4-c.6 | 1.189 | 0.841 | 1.627 | 0.615 | 1.186 | 0.843 |
| c.5-c.6 | 5.384 | 0.186 | 4.213 | 0.237 | 3.557 | 0.281 |
| a.1 | 0.200 | 5.010 | 0.386 | 2.590 | 0.200 | 5.000 |
| a.2 | 0.244 | 4.091 | 0.316 | 3.162 | 0.179 | 5.593 |
| a.3 | 0.218 | 4.583 | 0.319 | 3.130 | 0.281 | 3.557 |
| b | 0.151 | 6.640 | 0.325 | 3.080 | 0.189 | 5.278 |
| c.1 | 0.186 | 5.384 | 0.289 | 3.464 | 0.168 | 5.944 |
| c.2 | 0.167 | 6.000 | 0.291 | 3.440 | 0.243 | 4.121 |
| c.3 | 0.205 | 4.880 | 0.247 | 4.054 | 0.237 | 4.217 |
| c.4 | 0.253 | 3.956 | 0.386 | 2.590 | 0.179 | 5.593 |
| c.5 | 0.273 | 3.663 | 0.169 | 5.916 | 0.212 | 4.718 |
| c.6 | 0.150 | 6.654 | 0.312 | 3.201 | 0.199 | 5.013 |

## TABLE 15

## LOCAL PRIORITY WEIGHTS FOR ELEMENTS OF ITEM 5.2 HIERARCHY

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 0.701 | 0.550 | 0.218 | 0.080 | 0.455 | 0.753 | 0.200 | 0.234 | 0.333 | 0.519 | 0.644 |
| b | 0.202 | 0.210 | 0.067 | 0.685 | 0.091 | 0.063 | 0.400 | 0.080 | 0.333 | 0.177 | 0.242 |
| c | 0.097 | 0.240 | 0.715 | 0.234 | 0.455 | 0.184 | 0.400 | 0.685 | 0.333 | 0.304 | 0.114 |
| **Sum** | **1.000** | **1.000** | **1.000** | **0.999** | **1.001** | **1.000** | **1.000** | **0.999** | **0.999** | **1.000** | **1.000** |
| a.1 | 0.594 | 0.648 | 0.135 | 0.067 | 0.369 | 0.080 | 0.297 | 0.333 | 0.651 | 0.135 | 0.143 |
| a.2 | 0.249 | 0.122 | 0.584 | 0.218 | 0.182 | 0.685 | 0.163 | 0.333 | 0.223 | 0.281 | 0.429 |
| a.3 | 0.157 | 0.230 | 0.281 | 0.715 | 0.449 | 0.234 | 0.540 | 0.333 | 0.127 | 0.584 | 0.429 |
| **Sum** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **0.999** | **1.000** | **0.999** | **1.001** | **1.000** | **1.001** |
| c.1 | 0.160 | 0.308 | 0.309 | 0.232 | 0.167 | 0.563 | 0.263 | 0.068 | 0.215 | 0.069 | 0.089 |
| c.2 | 0.048 | 0.092 | 0.130 | 0.053 | 0.134 | 0.074 | 0.153 | 0.028 | 0.150 | 0.069 | 0.060 |
| c.3 | 0.029 | 0.099 | 0.075 | 0.504 | 0.374 | 0.085 | 0.146 | 0.424 | 0.137 | 0.087 | 0.379 |
| c.4 | 0.170 | 0.099 | 0.041 | 0.047 | 0.093 | 0.040 | 0.108 | 0.199 | 0.057 | 0.190 | 0.298 |
| c.5 | 0.501 | 0.296 | 0.407 | 0.129 | 0.162 | 0.207 | 0.276 | 0.252 | 0.327 | 0.423 | 0.127 |
| c.6 | 0.092 | 0.107 | 0.038 | 0.036 | 0.069 | 0.030 | 0.053 | 0.028 | 0.115 | 0.162 | 0.047 |
| **Sum** | **1.000** | **1.001** | **1.000** | **1.001** | **0.999** | **0.999** | **0.999** | **0.999** | **1.001** | **1.000** | **1.000** |

Results from QAF-USA Scoring. All of the QAF-USA scoring data is presented in the analysis section of this chapter.

Results from Feedback Questions. Appendix C contains 15 tables which present the detailed results from each of the 15 feedback questions. The questions fall into the four general categories of understandability, usability, believability, and applicability; and, they are labeled and numbered to match the analysis presented later in this chapter.

The analysis of this data relies primarily on histograms which show the frequency of the paired comparison responses; however, some of the individual scores are referenced during the analyses. Each table simply shows how each evaluator recorded his or her preference with respect to the question posed at the top of the table. For example, based on the responses to the question "Which method is easier to understand?" shown in Table 32, evaluator one had a *strong* preference for the QAF-USA method over the AHP-USA method as indicated by the "T" in the "5" column on the QAF side. The "T" indicates that evaluator one was *trained* in the QAF-USA method. A "U" indicates an *untrained* evaluator, meaning that the evaluator had not had formal QAF-USA training.

<u>Results from Elapsed Time Collection</u>. During the data collection for this study, each evaluator was asked to record beginning and ending times for the tasks involved with each of the scoring methods. Table 16 defines these tasks using the same terms that were used during data collection. For each of these tasks, with the exception of *Team Score Generation*, the beginning time was subtracted from the ending time to determine the elapsed time (in minutes) for the task. For *Team Score Generation* the elapsed time of Step 2.4 was increased by eight minutes to account for the nominal amount of time needed to enter four sets of individual scores into Expert Choice, synthesize a weight for the alternatives, and calculate the team percentage score.

## TABLE 16

## TASK DEFINITIONS

| Scoring Method | Task | Definition |
|---|---|---|
| AHP-USA | Step 2.1 | Evaluate level 1 (areas a, b, and c) of the hierarchy. |
| AHP-USA | Step 2.2 | Evaluate level 2 (sub-areas a.1, a.2, and a.3) of the hierarchy. |
| AHP-USA | Step 2.3 | Evaluate level 2 (sub-areas c.1 through c.6) of the hierarchy. |
| AHP-USA | Step 2.4 | Evaluate level 3 (the alternatives) of the hierarchy |
| AHP-USA | Team Score Generation | Generate a team score based on the geometric mean of four individual scores. |
| QAF-USA | Individual Scoring | Identify strengths and areas for improvement, then determine a percentage score for Item 5.2. |
| QAF-USA | Team Consensus Scoring | Discuss strengths and areas for improvement, then adjust individual scores to reach a team consensus score. |

The following tables present all of the elapsed times which are the basis for the time analysis histograms which can be found later in this chapter. The mean and standard deviation are calculated for each set of data only to give the reader a general feel for the location and variance of the data; these statistics are not used for any further analysis.

Table 17 summarizes the elapsed times for the tasks needed to evaluate the criteria (levels 1 and 2) of the Item 5.2 hierarchy during the AHP-USA process.

Table 18 summarizes the elapsed times for the tasks of evaluating the alternatives (actual and ideal) with respect to the leaves of the hierarchy, thus giving the base set of numbers for calculating the *Team Score Generation* elapsed times.

TABLE 17

ELAPSED TIMES FOR EVALUATING CRITERIA IN AHP-USA

| Eval # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean | StdDev |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Step 2.1 | 1 | 3 | 2 | 3 | 2 | 2 | 2 | 3 | 0.75 | 2 | 5 | 2.3 | 1.15 |
| Step 2.2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 3 | 2 | 1.7 | 0.65 |
| Step 2.3 | 3 | 5 | 3 | 3 | 4 | 4 | 3 | 3 | 2 | 4 | 7 | 3.7 | 1.35 |
| Total | 5 | 10 | 6 | 8 | 8 | 8 | 7 | 7 | 3.75 | 9 | 14 | 7.8 | 2.71 |

TABLE 18

ELAPSED TIMES FOR EVALUATING ALTERNATIVES IN AHP-USA

| Evaluator # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean | StdDev |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Step 2.4 | 10 | 40 | 4 | 3 | 7 | 10 | 8 | 1 | 7 | 7 | 10 | 9.7 | 10.47 |
| Team Score Generation | 18 | 48 | 12 | 11 | 15 | 18 | 16 | 9 | 15 | 15 | 18 | 17.7 | 10.47 |

Table 19 summarizes the elapsed times required for individual and team consensus scoring when the evaluators were using the QAF-USA method. Evaluator eight's individual scoring time is estimated based on the starting time for individual scoring and the starting time for team consensus scoring. This estimate was used because evaluator eight failed to record the ending time for his or her individual scoring effort. As shown, three other evaluators failed to record either start or stop times; thus, not even an estimated duration was possible. Also, there was some disagreement between the elapsed times for team C consensus scoring, therefore, the three times were averaged, resulting in the times shown in Table 20.

TABLE 19

**ELAPSED TIMES FOR INDIVIDUAL AND TEAM SCORING IN QAF-USA**

| **Evaluator #** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **Mean** | **StdDev** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Team | | A | | | | B | | | | C | | | |
| Individual | 9 | | 24 | 6 | 9 | 20 | | 27 * | 8 | | 25 | **16.0** | **8.82** |
| Team Consensus | 15 | 15 | 15 | 15 | 25 | 25 | 25 | 25 | 21 | 21 | 27 | **20.8** | **4.94** |

* Estimated from partial data

Table 20 summarizes the data used to generate the histograms later in this chapter. *Team Score Generation* times come directly from Table 18, while the *Team Consensus Score* times are the team-by-team averages of the observed elapsed times for the team consensus scoring tasks presented in Table 19.

TABLE 20

**ELAPSED TIMES REQUIRED TO GENERATE TEAM SCORES**

| Evaluator # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | **Mean** | **StdDev** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AHP-USA | 18 | 48 | 12 | 11 | 15 | 18 | 16 | 9 | 15 | 15 | 18 | **17.7** | **10.47** |
| QAF-USA | 15 | 15 | 15 | 15 | 25 | 25 | 25 | 25 | 23 | 23 | 23 | **20.8** | **4.69** |

At this point, all of the data has been presented. The remainder of this chapter will analyze this data using quantitative and qualitative techniques in order to answer the research questions that were posed in Chapter II.

## Analyses

This section presents the information needed to answer each research question. It is organized to match the order in which the research questions were posed.

1. <u>Is the proposed AHP-based USA scoring method feasible?</u> This question is answered by examining the accuracy of the proposed AHP-USA scoring method by comparing its scores to those generated by the traditional QAF-USA method, and by examining the ability of the AHP to adapt to the existing QAF criteria.

1.a. <u>Can the AHP-USA method generate an accurate aggregate score?</u> This question is answered by comparing individual, team, and historical scores using the RMS, MAD, and MAPE accuracy measures. Table 21 shows the individual percentage unit scores obtained using both methods and the difference between the scores.

TABLE 21

SUMMARY OF INDIVIDUAL SCORES

| Evaluator # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AHP Scores (%) | 13 | 17 | 15 | 45 | 15 | 30 | 25 | 40 | 17 | 27 | 21 |
| QAF Scores (%) | 20 | 30 | 30 | 40 | 20 | 20 | 70 | 70 | 10 | 45 | 40 |
| Difference | -7 | -13 | -15 | 5 | -5 | 10 | -45 | -30 | 7 | -18 | -19 |

Simply by looking at the differences between the numbers it appears that the two methods result in significantly different individual scores. This observation is supported when the significance of the RMS and the MAD are calculated as shown in Table 22.

TABLE 22

SIGNIFICANCE RATIOS FOR INDIVIDUAL SCORES

| RMS | MAD |
|---|---|
| 0.66 | 0.27 |

Recall that Saaty (1980:39) states that if the significance ratio of either the RMS or MAD is less than 0.1 then the two series of scores are nearly the same. Considering the resulting values, even the smallest value of 0.27 does not approach the threshold of 0.1. Therefore, the individual scores generated by both methods are not equivalent.

Equivalent RMS and MAD numbers were calculated for team scores as shown in Table 23 and Table 24. Again, the RMS and MAD significance ratios are much larger than 0.1. Therefore, the team scores generated by both methods are not equivalent.

**TABLE 23**

**SUMMARY OF TEAM SCORES**

| **Team** | **A** | **B** | **C** |
|---|---|---|---|
| AHP Scores (%) | 20 | 30 | 21 |
| QAF Scores (%) | 30 | 20 | 18 |
| Difference | -10 | 10 | 3 |

**TABLE 24**

**SIGNIFICANCE RATIOS FOR TEAM SCORES**

| **RMS** | **MAD** |
|---|---|
| 0.36 | 0.30 |

Finally, one more set of RMS and MAD significance ratios were calculated by comparing the set of AHP-based team scores with the historical score of 52.5 percent which was assigned when the unit's complete USA report was originally scored by a team of ASC evaluators. These results are shown in Table 26.

TABLE 25

SUMMARY OF AHP TEAM VS HISTORICAL SCORES

| Team | A | B | C |
|---|---|---|---|
| AHP Team Scores (%) | 20 | 30 | 21 |
| QAF Team Scores (%) | 30 | 20 | 18 |
| Historical Score (%) | 52.5 | 52.5 | 52.5 |
| AHP Difference | -32.5 | -22.5 | -31.5 |
| QAF Difference | -22.5 | -32.5 | -34.5 |

TABLE 26

SIGNIFICANCE RATIOS FOR TEAM SCORES

| | RMS | MAD |
|---|---|---|
| AHP Scores | 0.77 | 0.03 |
| QAF Scores | 0.81 | 0.05 |

In this case, the two accuracy measures dramatically disagree. Based on the RMS measure, both the AHP- and QAF-generated scores are significantly different than the historical unit score of 52.5 percent. But, the MAD measure indicates good agreement. Simply by observing the difference in the scores, it is easy to see that the MAD numbers should be disregarded in this case. However, another accuracy measure -- mean absolute percentage error (MAPE) -- can also be used to make a quantitative comparison of how closely the sets of scores match. To gain perspective on how MAPE compares to the RMS and MAD measures, the MAPE was calculated for individual, team, and historical scores. Table 27 summarizes the results of all three types of accuracy measures. Judging by the RMS and MAPE results, the team scores generated by *either* method are not equivalent to the historical unit score.

TABLE 27

SUMMARY OF SCORING ACCURACY

| | RMS | MAD | MAPE |
|---|---|---|---|
| **AHP vs. QAF - Individual Scores** | 0.66 | 0.27 | 43.7 |
| **AHP vs. QAF - Team Scores** | 0.36 | 0.30 | 33.3 |
| **AHP vs. Historical - Team Scores** | 0.77 | 0.03 | 54.9 |
| **QAF vs. Historical - Team Scores** | 0.81 | 0.05 | 56.8 |

In conclusion, the proposed AHP-USA scoring method did not generate scores which accurately matched individual, team, or historical scores which were derived by the QAF-USA method.

1.b. Can the AHP adapt to the existing QAF-USA criteria?

This question is answered by a subjective analysis of the task of deriving a hierarchy from the definition of QAF Item 5.2 and, by an analysis of the consistency ratios across the levels of the hierarchy. Specifically, the subjective analysis will discuss whether an appropriate hierarchy was constructed from the QAF evaluation criteria; while the consistency ratio analysis will provide insight into whether the evaluators were able to judge the relative importance of the elements in the hierarchy.

From a mechanical perspective, it was very easy to translate the QAF criteria into an AHP hierarchy. Because the QAF criteria are already well structured into *categories, items, areas,* and even more detailed elements (called *sub-areas* in this study), it took less than 30 minutes to create the hierarchy within the Expert Choice software from the narrative description of QAF Item 5.2. Most of this time simply involved entering the definitions for each of the criteria. Of course, many hours were spent preparing the AHP-USA briefing (Appendix A) and data collection package (Appendix B). However, for an actual implementation of the AHP-USA method, these tasks would be one-time events requiring only a single investment of time. So, this provides some insight into the ease with which the AHP can adapt to the QAF criteria; however, it does not answer the question of whether the resulting hierarchy is *appropriate*.

A hierarchy is appropriate if it satisfies the axioms underlying the AHP. Specifically, axiom 1 requires that decision makers be able to compare the elements of the hierarchy, axiom 2 requires that elements be relatively homogenous, axiom 3 requires that the importance of the criteria be independent of the properties of the alternatives, and axiom 4 requires that the hierarchy be sufficiently complete to address the problem. Axioms 1 and 2 are addressed later in the analysis of consistency ratios which follows the discussion of axioms 3 and 4 below.

The QAF criteria are independent of the alternatives because they are based on the Malcolm Baldrige National Quality Award criteria which are designed to be applicable across a broad range of businesses. Specifically, manufacturing, service, and small businesses are eligible to compete for the award. Thus, to be fair to this broad range of competitors, the criteria cannot be dependent on the units assessed.

By definition, the QAF criteria are complete for the purposes of unit self-assessment because they define the bounds of the USA process. Of course, only Item 5.2 out of the 28 total QAF items was used in this study. But, all of the areas and sub-areas within Item 5.2 were included in the AHP hierarchy. So, within the scope of this study, the Item 5.2 hierarchy was complete.

Based on these arguments, the Item 5.2 hierarchy meets the requirements for axioms 3 and 4. To determine whether axioms 1 and 2 are also met, and to answer whether the evaluators were able to consistently judge the criteria in the Item 5.2 hierarchy requires analysis of the paired comparison judgments. Specifically, we will look at the consistency ratios and the weights derived from the paired comparison judgments.

First, we will examine the consistency ratios for each evaluator at each level of the hierarchy. Figure 13 again shows the hierarchy derived from QAF Item 5.2. All 11 evaluators judged each level of this hierarchy using paired comparisons to determine the relative weight or importance of each element in the hierarchy.

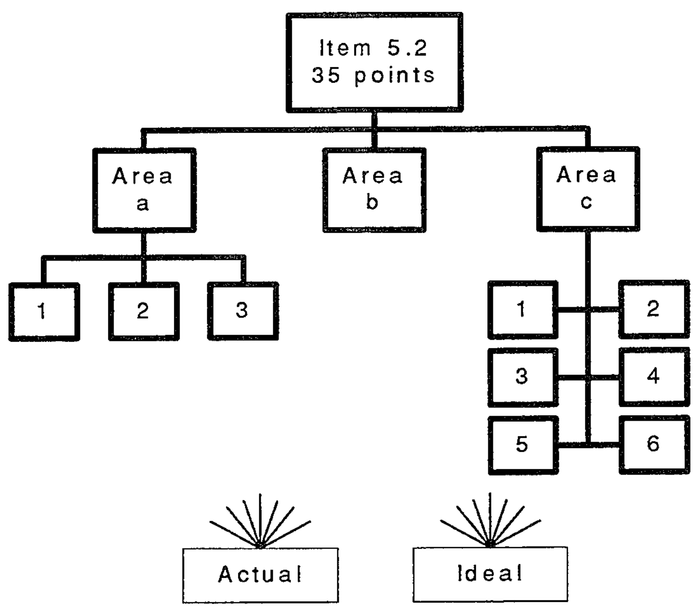



**Figure 13 Item 5.2 Hierarchy**

Recall that each level of the hierarchy requires one or more sets of paired comparison judgments. Table 28 defines the levels of the hierarchy and the set(s) of elements compared at each level.

TABLE 28

SETS OF PAIRED COMPARISONS FOR ITEM 5.2 HIERARCHY

| | Sets of Elements Compared | |
|---|---|---|
| Level 0 | None (single goal element for Item 5.2) | |
| Level 1 | Areas a, b, and c | |
| Level 2 | Sub-areas a.1, a.2, and a.3 | Sub-areas c.1, c.2, c.3, c.4, c.5, c.6 |
| Level 3 | Actual vs. Ideal with respect to each of the leaves (a.1, a.2, a.3, b, c.1 through c.6) | |

The paired comparison judgments for each of these sets of elements were entered into Expert Choice in order to calculate the consistency ratios shown in Table 29. The consistency ratios of all evaluators are presented for each level of the hierarchy.

TABLE 29

CONSISTENCY RATIOS

| | Evaluators | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Elements | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| a-c | 0.130 | 0.017 | 0.175 | 0.282 | 0.000 | 0.282 | 0.000 | 0.282 | 0.000 | 0.282 | 0.395 |
| a.1-a.3 | 0.051 | 0.004 | 0.130 | 0.175 | 0.833 | 0.282 | 0.245 | 0.000 | 0.282 | 0.130 | 0.000 |
| c.1-c.6 | 0.091 | 0.010 | 0.066 | 0.345 | 0.127 | 0.135 | 0.137 | 0.150 | 0.089 | 0.026 | 0.103 |
| a.1-b-c.6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Overall | 0.100 | 0.010 | 0.110 | 0.300 | 0.200 | 0.250 | 0.080 | 0.180 | 0.080 | 0.160 | 0.220 |

Note that for level three (a.1-b-c.6) consistency is perfect because only two alternatives (actual and ideal) are being compared, therefore inconsistency cannot exist. Finally, Table 29 also presents the overall consistency ratio for each evaluator's complete hierarchy.

Every bold number in Table 29 is over the desired 0.1 consistency threshold. There are three potential causes for overly large inconsistencies. Saaty (1980:92) identifies two of them when he summarizes that "greater inconsistency [than 0.1] indicates lack of information or lack of understanding." Forman (1993:23) identifies a third cause where failure to achieve an acceptable consistency ratio might be due to large disparities between the importance of the elements of the hierarchy (i.e., the elements too heterogeneous). For example, if element A is 3 times more important than element B, and B is 6 times more important than element C, then A should be 18 times more important than C. However, the AHP comparison scale only permits relative scores from one to nine. Therefore, the paired comparisons could accurately represent the relationships between A and B, and B and C, but not A and C. If this situation is recognized, then a consistency ratio somewhat greater than 0.1 is acceptable (Forman, 1993:23).

Exploring Forman's cause first, we need to look for signs of heterogeneity. Specifically, if one or more elements in the hierarchy were either excessively important or excessively unimportant, then they should be consistently weighted very high or very low by all of the evaluators. Unfortunately, we cannot blame heterogeneity for the widespread inconsistency in the judgments because only two elements even approach the extreme ends of the weighting scale. The histograms in Figures Figure 14, Figure 15, and Figure 16 show the frequency of local priority weights across ten intervals spanning the entire range of weights from zero to one. Only sub-areas "c.2" and "c.6" show any signs of consistently being rated extremely low. However, after examining the actual paired comparison judgments in Table 12, only 3 of the 99 judgments which involved either "c.2" or "c.6" used the extreme values (either 9 or 0.111) of the AHP priority scale. Therefore, the elements in the hierarchy appear to be sufficiently homogeneous, satisfying axiom 2.

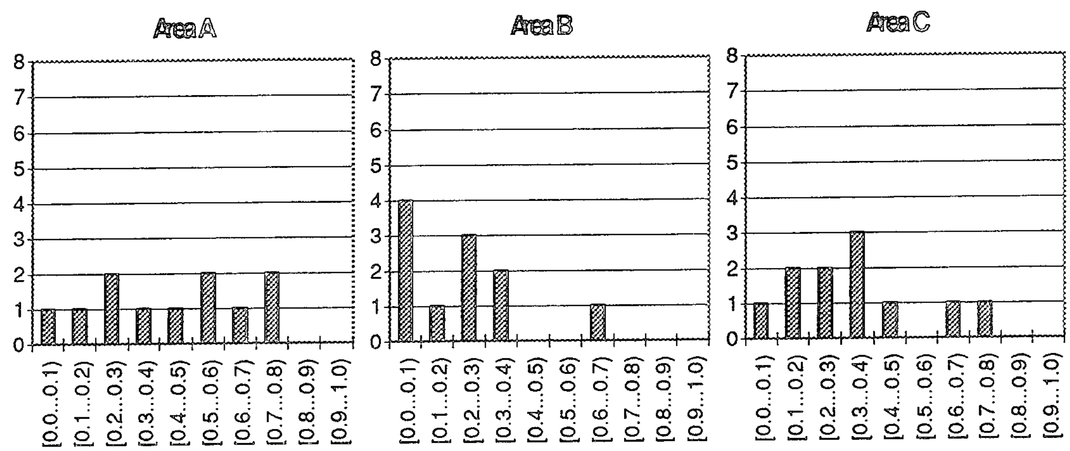



Figure 14 Histograms of Priority Weights for Areas A, B, and C

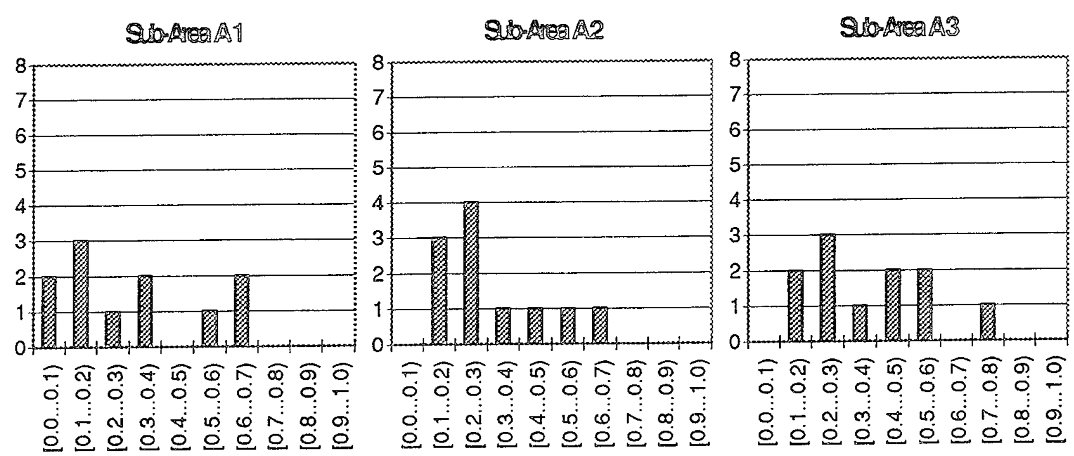



Figure 15 Histograms of Priority Weights for Sub-areas A.1, A.2, and A.3

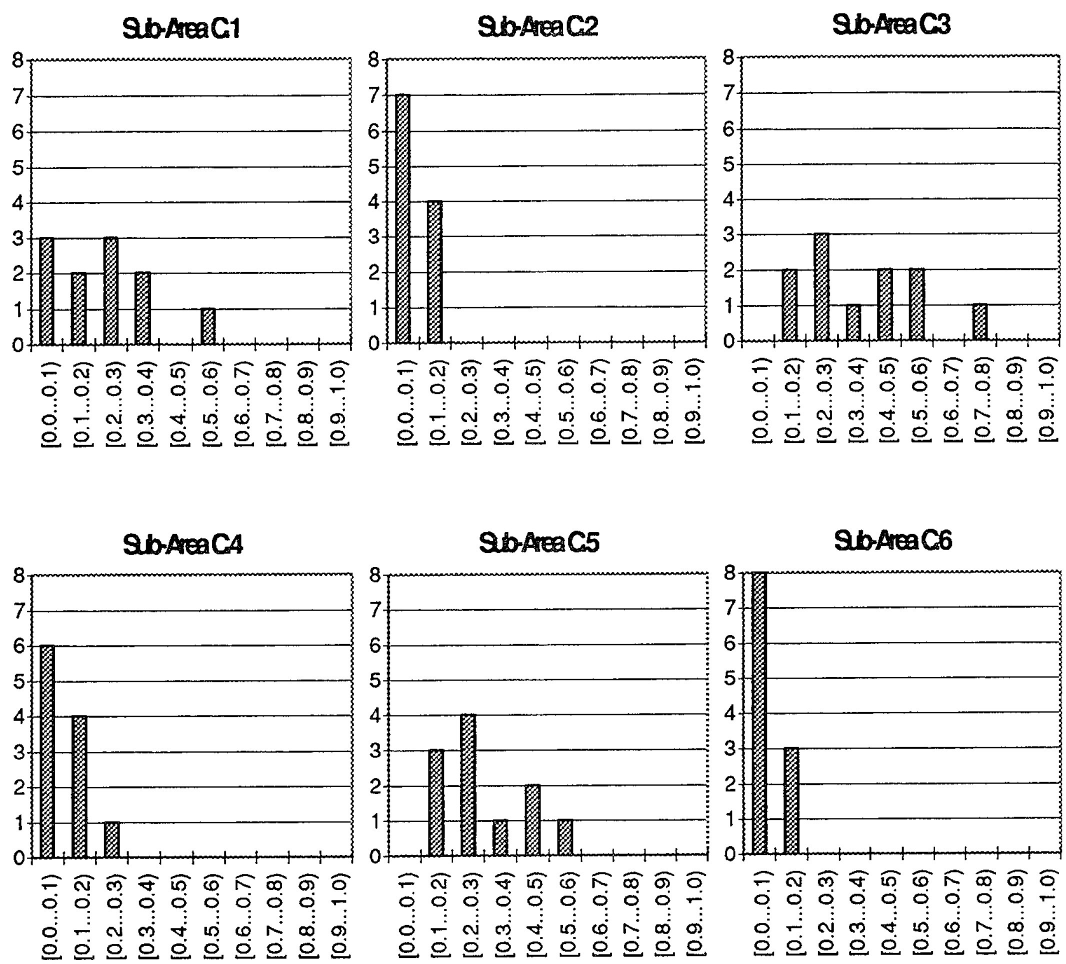



**Figure 16 Histograms of Priority Weights for Sub-areas C.1 through C.6**

This conclusion leaves us with one or both of Saaty's probable causes of inconsistency: 1) the evaluators were missing information, or 2) the evaluators did not clearly understand the judgments being made. It seems reasonable to surmise that those evaluators with the least training and experience would suffer from a lack of information about unit self-assessment. However, of the seven evaluators with *overall* consistency ratios greater than 0.1, only three (evaluators four, eight, and eleven) fell into the least trained/least experienced category, meaning that four of the most experienced evaluators were also inconsistent. Apparently a lack of USA information is not the sole cause of the observed inconsistencies; but, it is probably a strong contributing factor considering that only one (evaluator seven) of the four least trained/experienced individuals was able to achieve an acceptable overall consistency ratio.

Clearly, the ability to understand the judgment process is also a contributing factor. This ability can be affected by the three basic components of the AHP-USA scoring process: inputs, process, and evaluators.

First, are the formal inputs to the process. Namely, the USA report and the Item 5.2 hierarchy. The USA report can be eliminated as a cause of inconsistency because it played little, if any, role in the paired comparison judgments of the hierarchy *criteria*. In contrast, having a good understanding of the criteria in the hierarchy and their relative importance would have been vital to making sound, consistent judgments. The experienced evaluators should have had a good understanding of the criteria, but prior to this study the evaluators probably had not given much thought to their relative importance. And, the novice evaluators were surely at a disadvantage by having to learn and judge the criteria based on their written definitions alone. Therefore, lack of a clear understanding of the relative importance of the criteria is a likely cause of the observed inconsistency.

This apparent lack of understanding was surely exacerbated by the way the hierarchy was created and perhaps also by the nature of the QAF-USA criteria. While the hierarchy was easily derived from the QAF-USA criteria, this was not the typical way to create a hierarchy. Typically, the decision makers -- in this case the evaluators -- would create the hierarchy from scratch. In this fashion, the hierarchy becomes a custom framework which structures the overall problem into detailed decision criteria. When created from the top down, the evaluators have more time to become familiar with the meanings of the criteria in the hierarchy as they decompose the problem. This added familiarity should aid judgment consistency. Also, the wording and structure of the QAF-USA criteria may also have increased the potential for confusion. For example, when evaluating Item 5.2, Area *a*, for how a unit maintains the quality of its processes, the instructions for sub-area *a.1* asks the evaluators to look at a unit's "key processes and their requirements" while sub-area *a.2* focuses on "key indicators of quality and operational performance" (AFQI, 1993:21). It is easy to describe rational points of view that result in radically different judgments of these criteria. First, they could be considered equal because they are both fundamental aspects of process control. Second, *a.1* could be considered more important than *a.2*, because *a.1* is a prerequisite for *a.2*. Finally, *a.2* could be considered more important than *a.1*, because *a.1* is a more fundamental step, and therefore *a.2* demonstrates a more advanced step in a unit's quality program. Clearly, with 20 of 33 sets of criteria judgments being inconsistent as shown in Table 29, and the wide variety of weights shown in Figure 14 through Figure 16, the evaluators were largely unable to consistently judge the resulting hierarchy.

Despite the observed inconsistencies in the judgments, the strict requirements of axiom 1 are still met. Recall that axiom 1 requires that a decision maker (evaluator) be able to make comparisons between two elements (A and B) with respect to a single criteria (C) and express the strength of his or her preference. The evaluators in this study did make comparisons and did express preferences; however, they clearly had difficulty doing so in a consistent manner. So, does this indicate that the evaluators were *unable* to make comparisons and express preferences? No, it simply means that they were inconsistent. None of the axioms of the AHP require consistency; however, the results from the synthesis of inconsistent hierarchies must be used with caution.

At this point all of the axioms have been satisfied, but it is less clear whether the Item 5.2 hierarchy was truly *appropriate* for use with the AHP. Unfortunately, this question hinges on the true causes of the observed inconsistencies. This is unfortunate because there are many potential causes of inconsistency as previously discussed. Thus, without further research which takes stronger steps to control for these factors, the answer to the question of appropriateness becomes strictly a matter of subjective judgment. In this case, it appears the most likely causes of inconsistency were the evaluators inexperience with the AHP, and the inability to get immediate feedback regarding the consistency of their judgments. Computer-based training and data collection could help mitigate these problems.

In conclusion, we have explored the two facets of feasibility: accuracy and adaptability. The scoring results showed that the AHP-USA method did not produce scores which matched individual, team, or historical QAF-USA scores. Therefore, we conclude that the AHP-USA method is inaccurate if scores equivalent to the traditional QAF-USA method are desired. The answer regarding adaptability is less clear cut. It was shown that a hierarchy was easily constructed based on the QAF criteria which met all of the axiomatic requirements of the AHP. However, considering the inconsistency of the criteria judgments, many evaluators did not have a clear understanding supporting many of the judgments they made. The most likely causes of inconsistency were 1) inexperience with the AHP paired comparison process, and 2) unclear opinions regarding the relative importance of QAF criteria. Overall, this study shows that the AHP is not a feasible method for generating USA scores. However, this conclusion is based on the assumed requirement that the AHP-USA score match the QAF-USA score. As will be shown later, when the paired comparison judgments are consistent the AHP-USA method does show some promise at producing a range of scores which is smaller (more consistent) than the QAF-USA method. Therefore, if matching QAF-USA scores is less important than producing a more consistent score, then the AHP-USA method may still be considered a feasible approach.

2. Is the proposed AHP-based USA scoring method desirable? Having completed the feasibility assessment, we turn to the question of desirability. This question is answered by a set of six more specific questions:

a. Is the AHP method more consistent than the QAF method?

b. Is the AHP method more economical than the QAF method?
c. Is the AHP method more understandable than the QAF method?
d. Is the AHP method more usable than the QAF method?
e. Is the AHP method more believable than the QAF method?
f. Is the AHP method equally or more applicable to USA than the QAF method?

These questions are answered in the order listed in the following paragraphs.

2.a. Is the AHP method more consistent than the QAF method?

To answer this question, we examine the range of individual and team scores produced by both scoring methods. Table 30 summarizes the ranges of observed scores. The *all scores* columns show the range and difference (delta) for all of the observed individual and team scores. The *consistent scores* columns include the scores from only those evaluators that achieved an acceptable (less than 0.1) overall consistency ratio during the AHP-USA scoring. Finally, the highest and lowest scores have been thrown out for the *trimmed scores* columns.

TABLE 30

SCORING RANGES

| | All Scores | | | Consistent Scores | | | Trimmed Scores | | |
|---|---|---|---|---|---|---|---|---|---|
| | Min | Max | Delta | Min | Max | Delta | Min | Max | Delta |
| AHP-Individual | 13 | 45 | 32 | 13 | 25 | 12 | 15 | 40 | 25 |
| QAF-Individual | 10 | 70 | 60 | 10 | 70 | 60 | 20 | 45 | 25 |
| AHP-Team | 20 | 30 | 10 | | | | | | |
| QAF-Team | 18 | 30 | 12 | | | | | | |

Judging by all the scores, the AHP-USA process may be slightly more consistent than the QAF-USA process. Based on all of the individual scores, the AHP-USA process is significantly more consistent; but this advantage essentially vanishes when the team scores are considered. It is interesting that the AHP-USA's range for individual scores narrows significantly -- to a delta of 12, as shown in Table 30 -- if only the scores from the four most consistent evaluators are used. In sharp contrast, the range of QAF-USA scores from the same set of four evaluators did not narrow at all! These observations suggest that if consistency is maintained during the AHP-USA process, then the resulting AHP-based scores may be less sensitive to variability between evaluators. However, these observations might also be explained by evaluators having a central tendency when they are scoring the USA report with respect to the ten leaves of the hierarchy. Finally, when the extreme scores are eliminated from the individual scores of both processes, they are found to have essentially equal ranges. Therefore, since any USA scoring process would use teams, the AHP-USA process has little, if any, consistency advantage over the QAF-USA process.

2.b. Is the AHP method more economical than the QAF method?

The elapsed times gathered during the data collection will provide the answer to this question. Figure 17 shows the distribution of elapsed times observed for two tasks: 1) *Team Score Generation* (evaluating the USA report with respect to the leaves of the AHP hierarchy, plus eight minutes), and 2) *Team Consensus Scoring* (agreeing to a consensus team score using the QAF-USA method). Comparing these two tasks may seem like comparing apples to oranges because the *AHP-Leaves* times only consider the evaluators last step in the AHP-USA process (i.e., comparing alternatives with respect to leaves), while the *QAF-Team* times are taken from potentially lengthy consensus-building task.

However, there are several reasons why this is the fairest comparison that can be made with the available data.

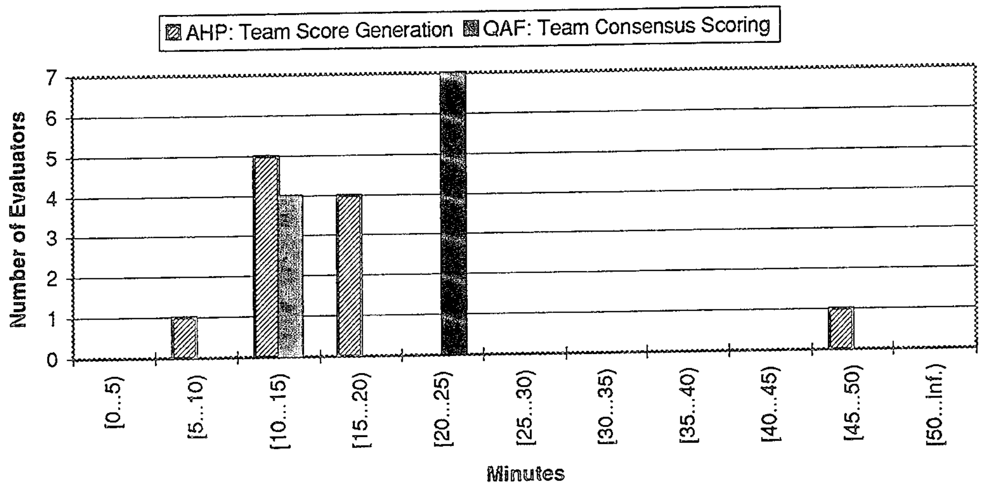



Figure 17 Histogram of Elapsed Times

First, it makes the most sense to compare those tasks which are necessary to generate the final team scores for the unit. If the AHP-USA method were implemented in the field, the full QAF hierarchy would be pre-constructed and pre-weighted, leaving only two tasks for the evaluators: 1) compare the alternatives (*actual* and *ideal*) with respect to every leaf in the hierarchy, and 2) enter the scores from the individual team members into an automated software package in order to generate a team score using the geometric mean. Within this study, data was collected for task one, and the duration for task two is approximately eight minutes for a hierarchy with 10 leaves and a team size of four evaluators. Thus, the observed elapsed times are each increased by eight minutes and then plotted on the histogram. Adding the eight minutes to each evaluators time tends to be a conservative bias because the task of data entry only requires *one* person for eight minutes rather than occupying four evaluators for an additional eight minutes.

Generating team scores with the QAF-USA process also demands a two step process: 1) individual scoring, and 2) team consensus scoring. During individual scoring, each evaluator generates a list of strengths and areas for improvement based on analysis of the USA report. This list and the QAF scoring guidelines are then used to determine a score for the unit. If the team members' individual scores vary by more than 20 percentage points then they must discuss their evaluations and agree to adjust their individual scores to bring the teams range of scores within a 20 point spread. In this study, both the individual scoring times and the team consensus scoring times were collected. However, only the team consensus times will be considered for this question because the evaluators strongly suggested that any AHP-USA scoring method include a step to generate a list of strengths and areas for improvement. So, if both methods were to include generation of equivalent lists, then both would require equivalent times, and there could be no efficiencies realized with either method.

Referring back to Figure 17, the AHP-USA method may have a slight chance of saving a little time; however, without more data to fill in the distributions more fully, no conclusions can be drawn from the data. Although, if a best case scenario is considered where the expected values for the AHP-USA and QAF-USA elapsed times would fall in the 10-15 and 20-25 minute ranges, respectively, would it be worth implementing the AHP-USA process assuming a 10 minute savings for each of the 28 items in a full QAF hierarchy? Four hours and forty minutes per scoring session are surely not enough savings to make a case for converting to the AHP-USA scoring method.

2.c. Is the AHP method more understandable than the QAF method?

To answer this question, the evaluators were asked to make paired comparison judgments between the two methods with respect to three questions:

1. Which method is easier to understand?
2. Which method is more intuitive?
3. Which method is simpler?

For these questions, the evaluators were asked to consider selected aspects of the two methods when making their judgments. For the AHP-USA method, the aspects of interest were the *hierarchy* and *paired comparisons*. For the QAF-USA method, the aspects were the *categories*, *items*, *areas*, *point values*, and *percentage scores* featured in this method. Figure 18 shows the evaluators' responses to these questions. The geometric mean is also plotted to show the central location of the paired comparison data.

As shown, the data indicates a slight to moderate preference for the AHP-USA method. The strongest responses in favor of the AHP-USA method were to the question of simplicity. However, during the data collection, one of the evaluators questioned the difference between "easier to understand" and "simpler." The phrase "simpler to use" helped clarify the difference. But, many of the evaluators had already answered the question as written. Thus, the responses to the third question may be mixed between the interpretations.

Another potential bias in favor of the AHP-USA method might have been caused due to the effects of the AHP-USA instruction on the four evaluators who had not had QAF-USA training. During the data collection process, significantly more time was spent explaining the AHP-USA method than was devoted to reviewing the QAF-USA method.

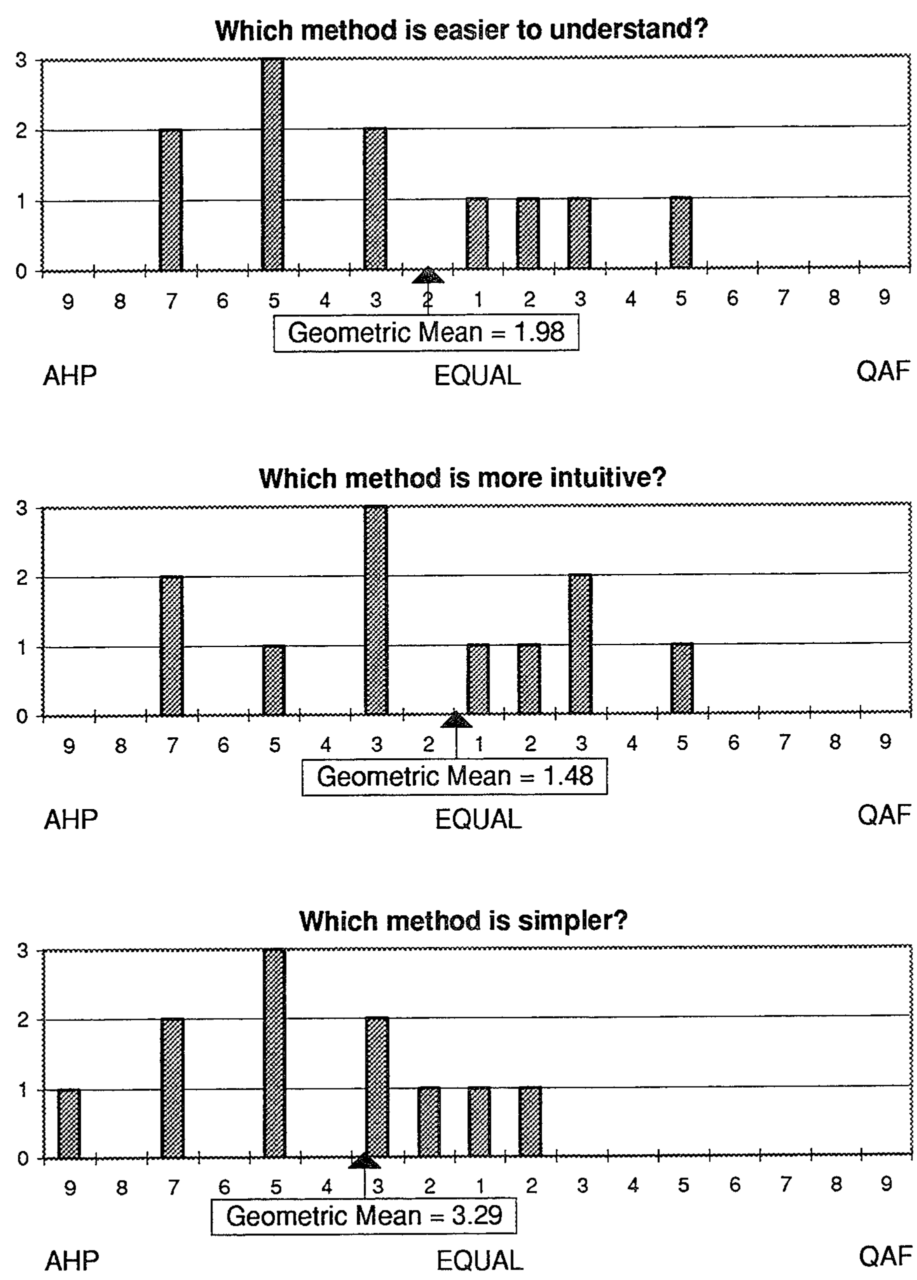



**Figure 18 Responses to Understandability Questions**

Thus, the four novice evaluators may have been biased towards the AHP-USA method, because of inadequate instruction in the QAF-USA method. Of course, if different levels of training and experience tend to bias the responses, then the other seven evaluators with a lot of QAF-USA training and experience should have been biased against the AHP-USA method. Considering these factors, the AHP-USA method can still claim a slight edge over the QAF-USA method with respect to understandability.

2.d. Is the AHP method more usable than the QAF method?

To answer this question the evaluators were asked to compare the two methods with respect to these questions:

1. For which method was it easier to determine a score for the unit?
2. Which method was easier to use overall in this scoring exercise?
3. Which method was a better tool to identify strengths and areas for improvement for the unit?

For question one, the evaluators were asked to focus on the task of performing paired comparisons for the alternatives with respect to the leaves of the AHP-USA hierarchy, compared to the tasks of individual and team consensus scoring used in the QAF-USA method. For question two, the focus for the AHP-USA method broadened to include the difficulty of judging the complete hierarchy, while the QAF-USA focus remained the same as question one. Finally, for question three, the AHP-USA focus returned to evaluating the alternatives, while the QAF-USA focus narrowed to the task of identifying strengths (pluses) and areas for improvement (minuses) during individual scoring.

Figure 19 shows the responses to the three questions. The responses to question one show a broad range of preference from *very strong* support for AHP-USA to *strong* preference for QAF-USA. Clearly, the group is split on which method was easier to use to determine a unit score. An identical range of responses was observed for question 2;

however, the geometric mean indicates a slight preference for the AHP-USA method. By examining the individual scores in Table 36, one may conclude that the novice evaluators uniformly preferred the AHP-USA method. Finally, question three elicited the strongest responses in favor of the QAF-USA method. The geometric mean indicates a *moderate* preference on average for the QAF-USA method as a way to identify strengths and areas for improvement. This is not a surprising response considering that the evaluators were not asked to use the AHP-USA method to identify strengths and areas for improvement. Instead, the evaluators used the standard QAF-USA worksheet to record strengths and areas for improvement, thus to answer question three, the evaluators had to compare actual use with hypothetical use.

Before leaving usability, it is worth noting that a common feedback comment from three of the evaluators dealt with the topic of identifying strengths and areas for improvement. Specifically, the evaluators suggested that a step be added to the AHP-USA method which would address the identification of narrative comments about the strengths and areas for improvement seen in the USA report. Considering that such narrative comments are particularly important as feedback to those units being assessed, any implementation of the AHP-USA process should include such a step. However, it should also be noted that the AHP-USA process provides a much more detailed set of numerical scores for each of the leaves (areas or sub-areas) in the hierarchy of QAF criteria. Such detailed numerical data could also help the units focus on exactly where they need to improve.

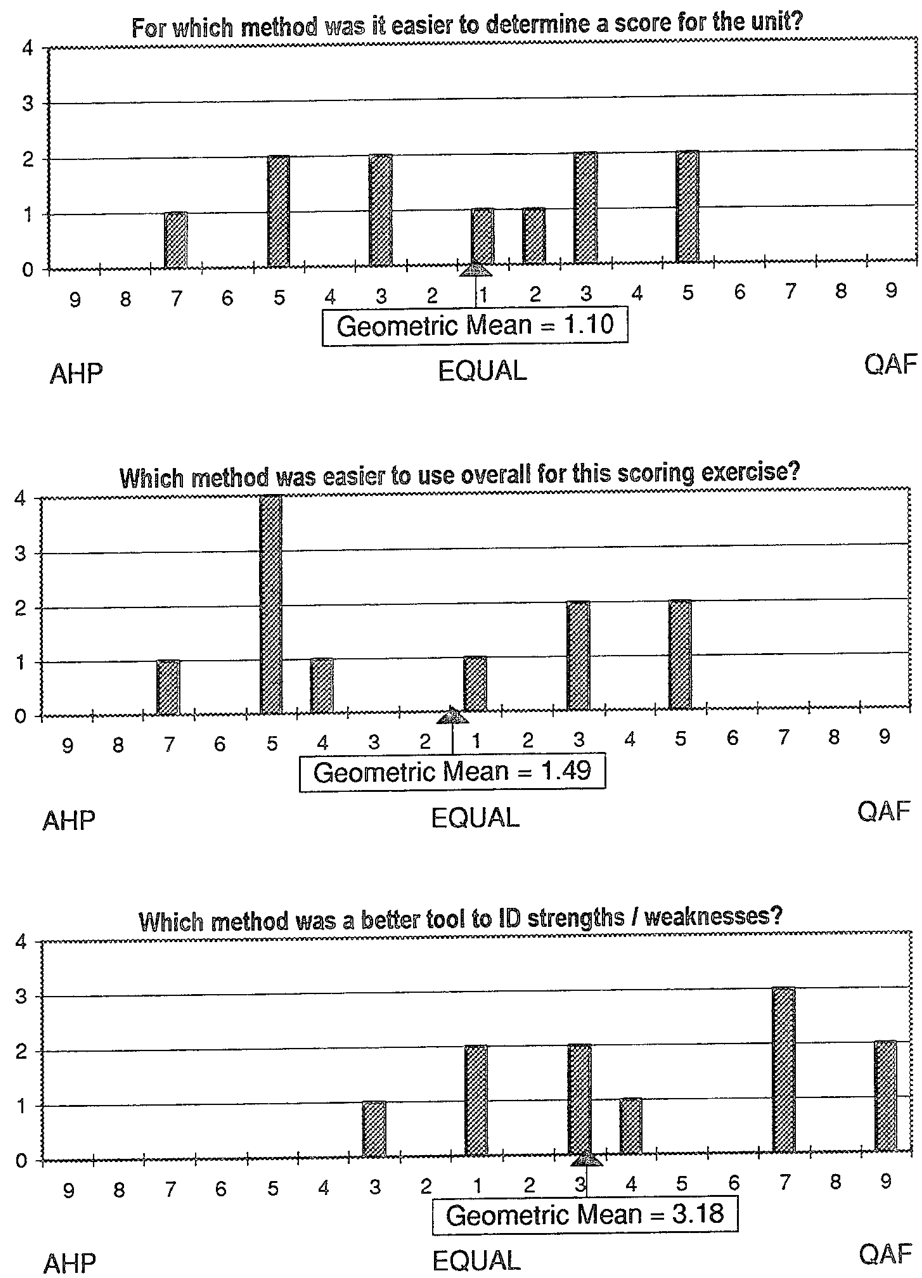



Figure 19 Responses to Usability Questions

2.e. Is the AHP method more believable than the QAF method?

To answer this question the evaluators were asked to compare the two methods with respect to the following questions:

1. If an "accurate score" is defined as "a score which truly represents a unit's performance relative to desired/planned performance," which method do you believe would produce the most accurate scores?
2. If a "consistent score" is defined as "a score which varies little between evaluators," which method do you believe would produce the most consistent scores?
3. Which method would you trust more to score unit self-assessments?

For these questions, no particular aspects of either method were used to help narrow the focus of the responses. In other words, the evaluators were asked to judge based on their overall impressions of both methods.

Figure 20 presents the responses to these questions. As shown, slight preferences for the AHP-USA method are indicated by the location of the geometric means for the questions about accuracy and consistency. The responses to the question of "trust" are very interesting (see Table 40). The geometric mean indicates a very slight preference for the QAF-USA method, but this indicator belies the strength of two evaluators' preferences. Specifically, the two most experienced evaluators expressed *very strong* and *extreme* preferences for the QAF-USA method. The novice evaluators split over this question: two of the four expressed a *moderate* preference for the AHP-USA method, one voted for *equality*, and one expressed a *strong* preference for the QAF-USA method.

In sum, the two methods appear to be equally believable with respect to their accuracy, consistency, and trustworthiness. However, the two most experienced evaluators clearly placed more trust in the QAF-USA method.

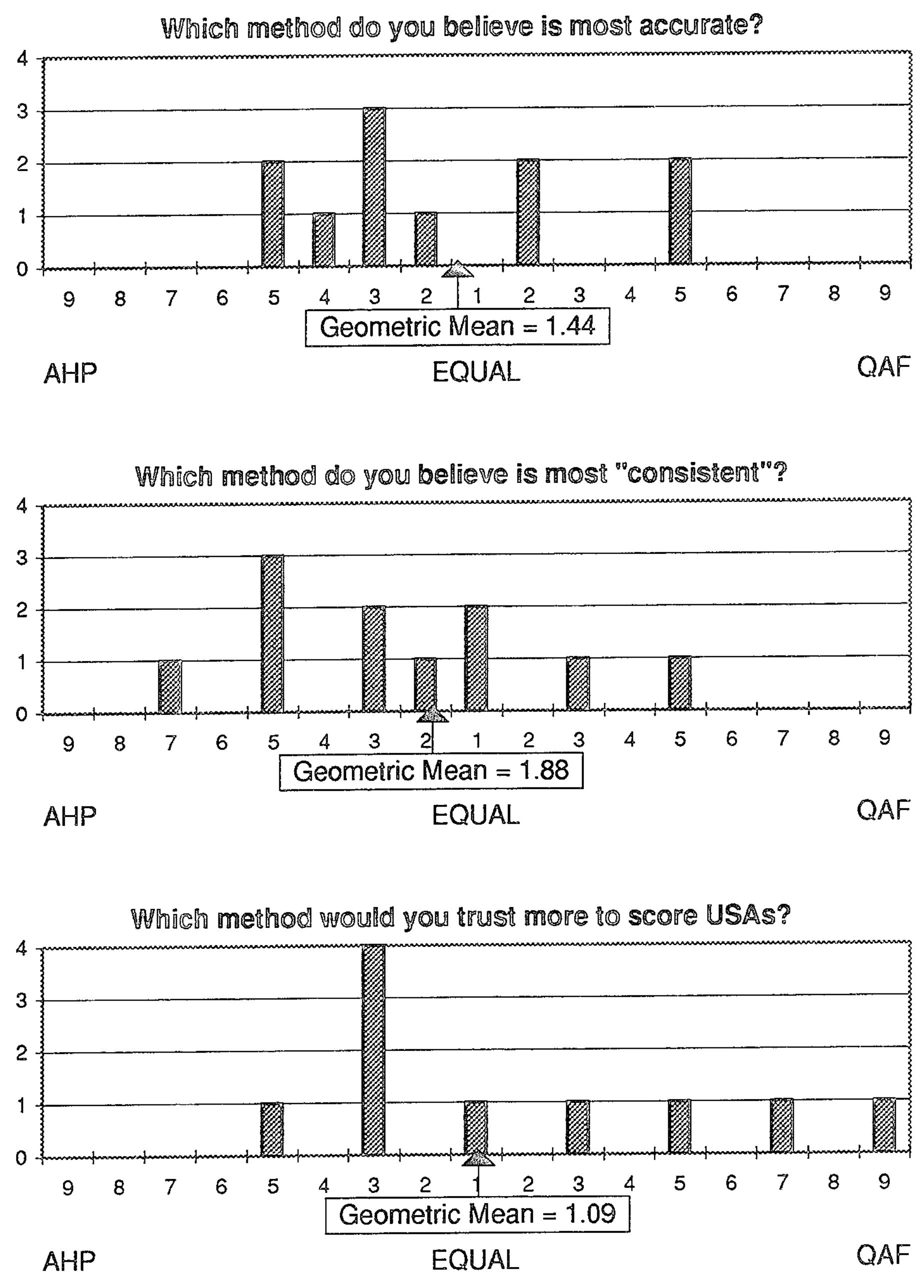



Figure 20 Responses to Believability Questions

2.f. Is the AHP method applicable to USA? To address this question, the evaluators were asked to compare the AHP-USA and QAF-USA methods with respect to six questions:

1. Which method would you prefer to use for future unit self-assessments?
2. Which method would you prefer to accurately show a trend in USA scores from year to year?
3. Which method would you prefer if USA were to be done by novice practitioners?
4. Which method would you prefer if USA were to be done by experienced practitioners?
5. Which method is best suited to a variety of USA practitioners (novice to experienced)?
6. Which method do you prefer overall?

As with the previous set of questions, no particular aspects of the methods were used to narrow the scope of the questions. Figure 21, Figure 22, and Figure 23 present the responses to these questions.

Questions one and six address overall preference for the scoring methods. As shown in Figure 21, the responses to both of these questions vary greatly. Looking at the individual judgments in Table 41 and Table 46, the four novice evaluators tended to favor the AHP-USA method, while the experienced evaluators were split. Thus, the geometric means for both sets of responses show slight preferences for the AHP-USA method. However, considering the variance in these sets of data, the only reasonable conclusion is that the two methods are equal in overall appeal.

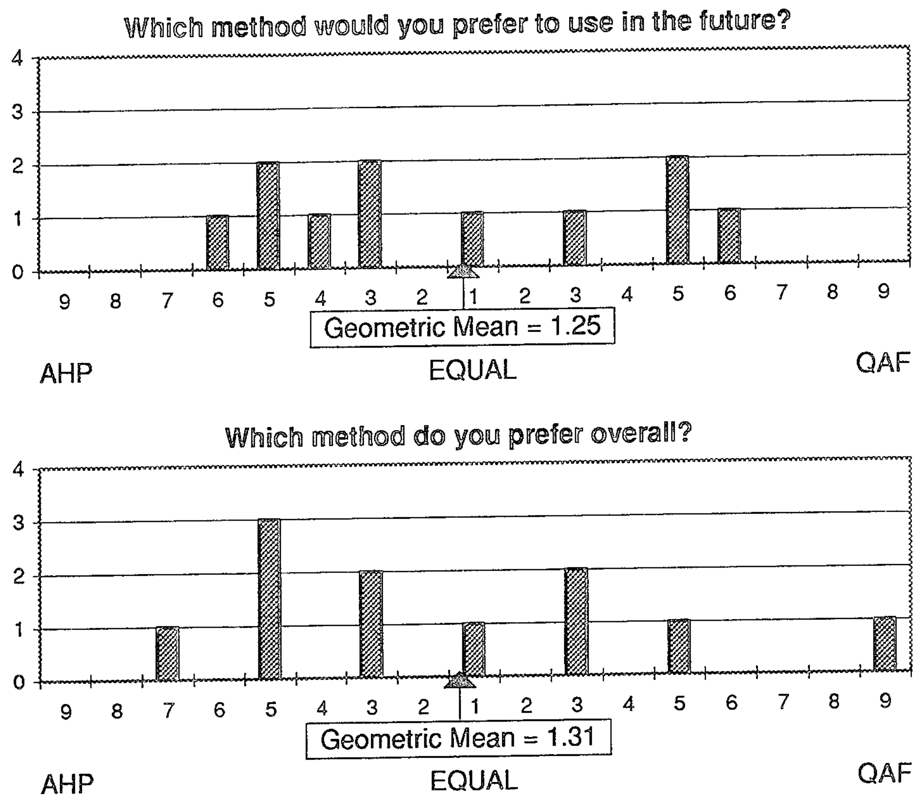



Figure 21 Responses to Applicability Questions 1 and 6

A similar conclusion can be drawn for question two, regarding the evaluators perceptions of either method's ability to show an accurate trend over time. As Figure 22 shows, the responses tend to slightly favor the AHP-USA method. However, the largest number of responses judged that the methods would be equal at showing a trend. This finding is consistent with the similar responses which were previously observed regarding accuracy and consistency. This is important, because both accuracy and consistency are needed to accurately show a trend.

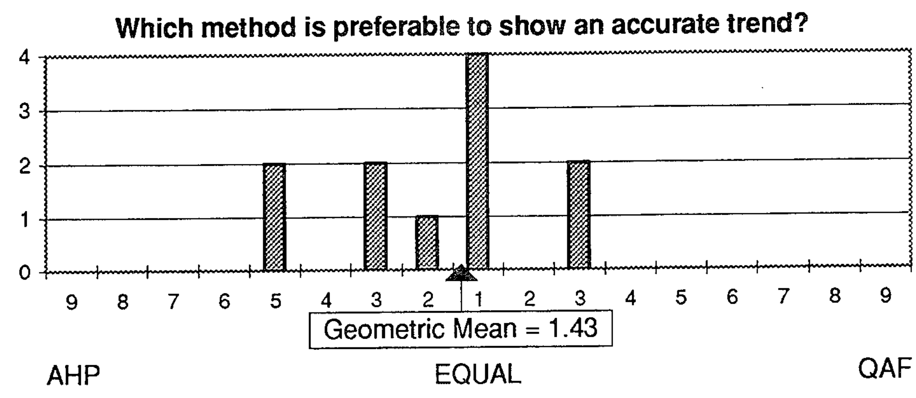



**Figure 22 Responses to Applicability Question 2**

The most definitive set of responses came from questions three, four, and five, as shown in Figure 23. This series compared methods by considering which scoring process would be best if the users were novices, experienced hands, or a mix of both. For question three, the geometric mean indicates a *moderate* preference for the AHP-USA method if novices were doing the scoring. And, the novice evaluators all expressed *moderate* to *very strong* preference for the AHP-USA method as shown in Table 43. In contrast, the responses to question four show only a slight preference for the QAF-USA method if used by experienced evaluators. Yet, three out of the four novice evaluators joined two of the experienced evaluators in expressing *very strong* to *extreme* preference for the QAF-USA method under these circumstances, while the rest of the experienced evaluators tended to favor the AHP-USA method (see Table 44). Based on this insight, the geometric mean probably understates the preference for the QAF-USA method with regard to question four. Finally, with question five, the pendulum swings back towards the AHP-USA method, where the geometric mean indicates a slight preference. In this case, most of the novice and experienced evaluators expressed similar degrees of preference (see Table 45).

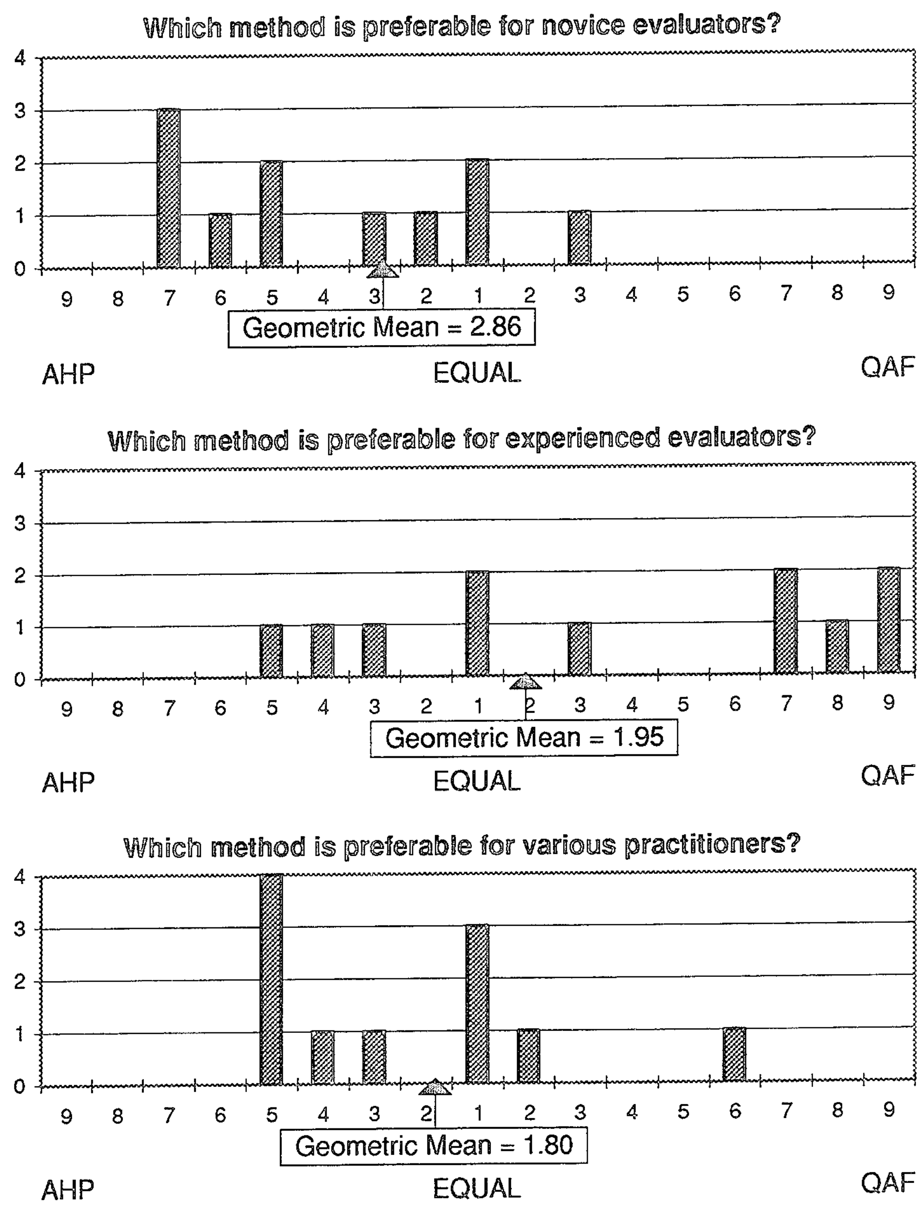



Figure 23 Responses to Applicability Questions 3, 4, and 5

# V. Conclusions and Recommendations

This chapter provides an overall review of the study and then presents some broad conclusions based on the findings in Chapter IV. Also, recommendations for potential implementation and further research are offered.

## Review

Chapter I began with a statement of the objective of this research, which was to examine the feasibility and desirability of using the analytic hierarchy process (AHP) to aggregate organizational performance metrics. Then, the chapter introduced the need for comprehensive performance measurement as a basis for effective management control. We saw that an aggregate measure, formed from several individual metrics, could help track the progress of a complex, multidimensional project. The Air Force's efforts, to use a unit self-assessment (USA) process for evaluating progress towards Total Quality Management (TQM) practices, were introduced as an example of where an aggregate measure was already being used to evaluate a complex process. Finally, the weaknesses in the USA process, namely, inconsistency and poor economy, were introduced.

Chapter II reviewed some of the available literature regarding performance measurement and aggregation. Specifically, the literature showed that performance measurement is vital for effective control because it provides the necessary framework for gauging progress towards objectives. After establishing the need for performance measurement, it was shown that a performance metric consists of *an operational* *definition*, *a measurement* over time, and a method of *presentation* which communicates vital information about processes and activities to decision makers.

With this definition in mind, Chapter II then showed that when objective, quantitative measurements are available (e.g., bushels of wheat harvested and gallons of diesel fuel consumed), one must still consider whether to form a *static* or *dynamic* measure of productivity. A static measure is used to show a *snapshot* of performance at a particular point in time. A dynamic measure, on the other hand, is used to show change from one period to the next. These productivity measurements (output/input ratios) can be classified by the number of input *classes* (labor, capital, energy, data, materials) included in the denominator of the measure. A *partial-factor* measure only includes one of the five classes, while *multifactor* and *total-factor* measures include *several* or *all* of the classes, respectively. In essence, multifactor and total-factor measures are a simple way to combine several diverse measurements into a single aggregate measure. By its name, a *total*-factor measure suggests that it can address the totality of an organizations measurement needs. Alas, such is not the case, because all organizations inherently have many dimensions that do not make sense to measure from a productivity -- outputs over inputs -- viewpoint.

Instead, more comprehensive measurement strategies must be used in order to accurately capture the overall performance of an organization. Specifically, we briefly defined the two basic *centralized* (top-down) and *decentralized* (bottom-up) measurement approaches, with a third approach defined as a hybrid the two. Also, the importance of multidimensional measures was driven home by considering the diverse statistics needed to measure the performance of a baseball team, or the many vital indicators that help a doctor assess the health of a patient. Clearly, it is important to use a variety of measures to gain a more accurate picture of organizational performance; but, when an activity is very complex (e.g., the mission of a corporate division), or difficult to measure in an objective quantitative way (e.g., implementing TQM), then it becomes difficult to draw meaning from numerous, separate performance metrics. Under these conditions, an aggregate measure is needed to reduce or simplify the data for use by high-level decision makers.

Unfortunately, it can be challenging to create a suitable aggregate metric. One challenge is posed by the relative importance of different measures; rarely will they be of equal importance with respect to indicating overall performance. A second challenge arises from the different scales associated with each detailed metric. For example, how do you combine a count of a baseball team's *unforced errors* with a count of *runs batted in*? The ideal number of errors is zero, while the ideal number of runs is arguably infinite. These factors make it difficult to combine metrics into a single aggregate measure.

At this point the discussion shifted gears from the general topic of performance measurement to the more specific function of aggregation. We defined a model of aggregation which to help compare and understand different aggregation methods. Then six different methods which had the potential to aggregate performance metrics were presented. One of these methods was the QAF-USA scoring method itself, the other five came from the review of performance measurement literature. Of these methods, the AHP was selected for further exploration based primarily on its logical structure, ease of use, and potential to be an improvement over the QAF-USA scoring method.

Because of its selection, we provided a more detailed description of the AHP which discussed its origins, applications, and definition. Specifically, we showed that the AHP was developed in the early 1970s by Thomas Saaty to address military contingency planning. As discussed, it has been widely studied and applied within the multicriterion decision-making arena, but its application to organizational performance metrics has not been explored. Nevertheless, its ability to derive preference weightings from subjective paired comparison judgments seemed well suited to the unit self-assessment task.

After the detailed review of the AHP, we looked at the unit self-assessment process. Specifically, we showed how the USA process adopted the 1993 version of the Malcolm Baldrige National Quality Award criteria. Thus, the USA criteria are structured into a three tier framework (or hierarchy) consisting of 7 Categories, 28 Items, and 92 Areas. Each of the 28 Items is classified as either an *approach/deployment-* or *results-*oriented item. From this complete structure, we chose Item 5.2 (Process Management: Product and Service Production and Delivery Processes) as a suitable focus for this study.

Finally, Chapter II listed the specific research questions for this study. These questions were designed to address the *feasibility* and *desirability* of using the AHP as a replacement for the QAF-USA scoring process. Under these broad questions, more detailed research questions considered the consistency, economy, understandability, usability, believability, and applicability of the proposed AHP-USA method to the task of unit self-assessment. Specifically, the research questions were:

1. Is the proposed AHP-based USA scoring method feasible?
    a. Can the AHP-USA method generate an accurate aggregate score?
    b. Can the AHP adapt to the existing QAF-USA criteria?

2. Is the proposed AHP-based USA scoring method desirable?
    a. Is the AHP method more consistent than the QAF method?
    b. Is the AHP method more economical than the QAF method?
    c. Is the AHP method more understandable than the QAF method?
    d. Is the AHP method more usable than the QAF method?
    e. Is the AHP method more believable than the QAF method?
    f. Is the AHP method equally or more applicable to USA than the QAF method?

Chapter III presented the methodology used in this research. Specifically, it presented the simple experimental approach taken to evaluate the proposed AHP-based USA scoring process by comparing it to the traditional Quality Air Force (QAF) USA method. In this approach, a group of 11 evaluators scored Item 5.2 based on an excerpt from a 1994 unit self-assessment report. The report had been written by a unit within the Aeronautical Systems Command and had been scored, at time of submittal, in accordance with the usual QAF-USA process. The experimental scoring for this study was done using both the AHP-USA and QAF-USA methods, and 15 feedback questions were asked to aid the comparison of the methods.

The resulting scores, AHP consistency ratios, elapsed times, and feedback results were assessed using a variety of analysis tools. Three quantitative accuracy measures (RMS, MAD, and MAPE) were used to determine whether both methods produced equivalent individual and team scores. Objective assessments of the consistency of the evaluators' judgments were based on simple observations of the consistency ratios from the AHP. Histograms were used to assess the observed elapsed times and the responses to the feedback questions. Also, some purely qualitative assessments were made regarding the ability of the AHP to adapt to the QAF-USA criteria.

Chapter IV presented the results, analysis, and detailed findings of this research. A brief review of these findings will set the stage for the broader conclusions and recommendations which follow. This review is organized by addressing each of the research questions in turn.

First, we look at the two research questions designed to address the broader question of feasibility. Specifically, we asked whether the AHP could 1) generate an accurate aggregate score, and 2) adapt to the existing QAF-USA criteria. To answer the first question we compared the individual, team, and historical (original ASC-generated) scores from both methods. The intent of the comparison was to determine whether the AHP-USA scores closely matched the equivalent scores from the QAF-USA scores which were generated either during the scoring experiment or earlier during the original USA scoring performed by ASC personnel. A close match is indicated by RMS or MAD significance ratios which are less than 0.1, or MAPE values which are less than 10. As indicated by the large values in Table 31, these accuracy measures show that the AHP-USA method does not generate aggregate scores which are accurate (equivalent) when compared to the QAF-USA scores.

TABLE 31

SUMMARY OF SCORING ACCURACY

| | RMS | MAD | MAPE |
|---|---|---|---|
| **AHP vs. QAF - Individual Scores** | 0.66 | 0.27 | 43.7 |
| **AHP vs. QAF - Team Scores** | 0.36 | 0.30 | 33.3 |
| **AHP vs. Historical - Team Scores** | 0.77 | 0.03 | 54.9 |
| **QAF vs. Historical - Team Scores** | 0.81 | 0.05 | 56.8 |

To answer the second question we discussed whether an appropriate AHP hierarchy was constructed from the QAF criteria, and whether the evaluators were able to consistently judge the resulting hierarchy. To construct a hierarchy from existing criteria required being able to structure the criteria into a hierarchical form which meets the four axioms of the theory behind the AHP. From a mechanical perspective, it was very easy to create a hierarchy from the QAF criteria because they were already structured in an outline or hierarchical form. And, we showed that this hierarchy did meet the four axioms of the AHP theory. However, based on the widespread inconsistency of the evaluators judgments, it is not clear whether the hierarchy was truly *appropriate* for use with the AHP. A lack of experience with the AHP and the inability to get immediate feedback with regard to the consistency of paired comparison judgments are the most probable causes of the observed inconsistencies. But, without additional research, it is impossible to say whether the AHP can truly adapt to the QAF-USA criteria.

Second, we look at the six questions designed to address the issue of desirability. Specifically, we looked at 1) the consistency of the scores from both methods, 2) the economy of both methods, 3) the relative understandability of both methods, 4) the relative usability of both methods, 5) the relative believability of both methods, and 6) the applicability of the AHP to the USA task.

The observed ranges of individual and team scores from both methods suggest that the AHP-USA method has a slight potential to be more consistent than the QAF-USA method. Specifically, as shown in Table 30, the AHP-USA method generated individual scores which spanned a range only half as wide as the QAF-USA scores. Thus, based on individual scores alone, the AHP-USA method was more consistent than the QAF-USA method. However, this consistency advantage essentially vanished when the ranges of the team scores were compared. Also, since this study only provides a single data point for observing the variance of scores, it is impossible to say whether the consistencies observed in the individual scores would hold up over repeated applications.

Similarly, little, if any, economic advantage was observed for the AHP-USA method. At best, the AHP-USA method may save approximately 10 minutes for each of the 28 QAF evaluation items, thus saving an average of four hours and forty minutes for each scoring session. Even given this optimistic scenario, the AHP-USA does not provide a significant economic advantage over the current QAF-USA method.

Regarding understandability, the AHP-USA method had a slight advantage over the QAF-USA method. In particular, the evaluators strongly favored the AHP-USA method when they compared the methods based on their perceived *simplicity*.

Regarding usability, both methods were equally preferred, except for the task of generating narrative *strengths and areas for improvement* where the QAF-USA method was strongly preferred. At first glance, the equality here is somewhat surprising in that the paired comparison judgments used in the AHP are a significantly different way to score a unit when compared to the conventional QAF-USA method. However, when one looks at how individual evaluators compared the methods, the evaluators without prior QAF-USA training tended to prefer the AHP-USA method. Therefore, the equality in usability comes primarily from the balance in the composition of the evaluation team.

Regarding believability, both methods were judged essentially equal. Specifically, the AHP-USA method may hold a slight edge over the QAF-USA method for both its *perceived* accuracy and consistency; however, the evaluators split over which method to *trust* with the task of USA scoring. And, two of the most experienced evaluators expressed extreme preference for the QAF-USA method on the question of trust. So, overall neither method is significantly more believable than the other.

Similarly, regarding overall applicability to unit self-assessment, both methods were equally preferred. When asked which method they would rather use in the future and which they preferred overall, the evaluators were widely split between the methods. When asked which method would best show a trend over time, the evaluators were again divided. This time with a plurality of evaluators judging the methods equal in ability to show a trend. Finally, when asked to choose which method would be best for novice evaluators, experienced evaluators, or a mix of both, the AHP-USA was favored for novices and mixed groups, while the QAF-USA method was favored for experienced practitioners. Overall, neither method held sway over applicability.

## Conclusions

Based on the results of this study, the proposed AHP-USA scoring method appears to be neither feasible nor desirable as a replacement for the current QAF-USA scoring method. From a feasibility perspective, the analysis showed that the AHP-USA method did not produce equivalent scores for either individual or team scoring efforts. In particular, the AHP-USA method generated predominately lower scores than the individual scores generated using the QAF-USA method. The team scores from both methods were in closer agreement, but not to the desired degree. Also, when the judgment consistency of the evaluators was examined, through the use of the AHP's consistency ratio, many of the judgments (which drive the AHP results) were shown to be overly inconsistent. With additional AHP training and an automated AHP implementation using commercial software, the inconsistency problem could be addressed. And, the analysis shows that the spread of scores narrows (i.e., becomes more consistent) with better judgment consistency; but, this does not correct the differences between the AHP- and QAF-based scores. Therefore, the proposed AHP-USA method is not a feasible *drop-in* replacement for the QAF-USA scoring method. In other words, if equivalent scores are deemed more important than consistent scores, then the proposed AHP-USA process is not a feasible option. However, if score consistency is more important than matching the traditional QAF-USA scores, then the AHP-USA scoring process should not be written off as unfeasible.

From a desirability perspective, there was no clear winner between the AHP-USA and QAF-USA methods. As mentioned, the AHP-USA method showed some potential for producing more consistent scores than the QAF-USA. Consistency is a desirable attribute; however, the potential improvement in consistency is marginal if only team scores are compared. Regarding the economy of the two methods, no advantage could be seen for the AHP-USA method. However, the data indicates that the evaluators tended to think the AHP-USA method was slightly more understandable than the QAF-USA method. In particular, the AHP-USA method was considered a *simpler* process. This could mean less training time for future evaluators, but this is not certain. With the exception of generating *strengths and areas for improvement* comments, the two methods were equally usable. The QAF-USA method was preferred for generating narrative feedback to the units with regard to their strengths and areas for improvement. However, the important conclusion from this preference is that narrative feedback is an important part of the USA process, and should be included in either method. If a narrative feedback step were added to the AHP-USA method, it could be as effective as, and perhaps even preferred over, the QAF-USA approach. Also, we saw that both methods were equally believable, however two of the most experience evaluators strongly preferred the QAF-USA method when asked which method they would trust more as a USA scoring method. For overall applicability, both methods were equally preferred. However, a preference was shown for the AHP-USA method if the USA practitioners were either relative novices, or a mix of novice and experienced evaluators. If the USA practitioners were all experienced, then the QAF-USA method was preferred. On balance, neither method is clearly more desirable than the other.

## Implementation Recommendations

Based on the results of this study, the proposed AHP-USA method should not be used to replace the existing QAF-USA scoring process. As previously discussed, the AHP-USA method is not a feasible approach for generating scores which are equivalent to those generated using the current QAF-USA process. Also, the AHP-USA is not significantly more desirable than the QAF-USA method in any particular aspect. Therefore, the AHP-USA method shows no clear advantage over the current QAF-USA process.

If a trial implementation of the AHP-USA method is desired despite the preceding recommendation, then several changes should be considered to improve the process. First, measures must be taken to improve the consistency of the paired comparison judgments. Specifically, evaluators should be provided with three to five hours of hands-on AHP training. Microcomputers with AHP software would greatly facilitate this training and improve the judgment of the evaluators by giving immediate feedback about the degree of consistency as every judgment is made. Furthermore, consistency may improve if each evaluator rank orders the criteria in question before making the paired comparison judgments. With the rank order at hand (and in mind), it should be easier to judge the relative preference of one criterion over another. These improvements should improve individual consistency and, in turn, the overall consistency of the AHP-USA method.

Second, the evaluation of the *actual* and *ideal* alternatives should be made as objective as possible. There are a couple of ways that the objectivity of the evaluations could be improved. One approach would involve providing *example attributes* for each of the lowest level criteria (the leaves) in the AHP-USA hierarchy. A couple of publications from the Federal Quality Institute (1993; 1995:57-85) may help develop such examples. In turn, these example attributes would help define the meaning of each leaf criteria and give the evaluator a more concrete understanding of what sort of unit behaviors (attributes) to look for in the USA report. Thus, when an evaluator is looking for "key processes and their requirements," as required by sub-area a.1 of Item 5.2, he or she will be able to look at specific world class examples of *processes* and *requirements* to get a better idea of what an *ideal* unit might use for this criteria. Then, a more objective comparison of the *actual* unit performance (based on observable attributes in the USA report) can be made with respect to the theoretical *ideal* performance (as defined by the example attributes). Another improvement to the AHP-USA method might be had by using *absolute* measurements and the *ideal mode* of synthesis as described in *recommendations for further research* section below. The absolute measurement mode requires a pre-defined scale which, like the example attributes just mentioned, should help clarify the meaning and application of each evaluation criterion.

## Recommendations for Further Research

Researching the feasibility and desirability of using a modified version of the AHP-USA scoring process for USA may be worthwhile. Specifically, the AHP can be structured to use *absolute* measurements from a pre-defined scale to assess alternatives. This means a unit would be evaluated based on a pre-defined scale rather than using paired comparisons to judge *actual* versus *ideal* performance on a relative basis. Also, another modification, even if relative comparisons are retained, would be to use the *ideal mode* of synthesis. This should affect the resulting scores generated using the AHP, and may improve the AHP-USA's performance with regard to matching the QAF-USA scores.

Researching the feasibility and desirability of using the MCP/PMT aggregation method for USA may have value. This method was judged in Chapter II as having the next best potential for addressing the USA deficiencies. Thus, it would be a logical choice for further research. Because the MCP/PMT is similar to the QAF-USA method, gains in economy are unlikely. However, if an MCP/PMT approach is used with a detailed level (i.e., equivalent to the sub-areas used in this study) of criteria, then the MCP/PMT approach has potential to improve the consistency of the USA scoring process.

Repeating this study might also be of value as a precursor to any implementation. However, the weaknesses encountered during this study should be addressed as part of any further study. First, additional time should be dedicated to training the evaluators on the AHP-USA scoring method. Whether trained or untrained in USA, the evaluators' judgment consistency would improve if a full hour of hands-on practice -- perhaps using the tennis ball selection example -- is provided before attempting the AHP-USA scoring. Ideally, the practice, and data collection, should use personal computer-based software to provide immediate feedback regarding judgment consistency. Also, for examiners who are untrained in the USA process, a minimum of two hours should be dedicated to providing an overview of the unit self-assessment process, with an emphasis on interpreting the QAF-USA scoring guidelines. This emphasis could take the form of providing world-class *example attributes* (examples of behavior) which, as previously discussed, may improve the objectivity of AHP-USA scoring as well. Second, as with most studies, a larger sample size would provide more confidence in the results. In this case, while it would be unwieldy to involve a significantly larger group in a single training and data collection session, smaller groups -- perhaps three to five people -- could be involved in many separate sessions. Of course, the benefits of being able to draw statistical inferences from perhaps 90 to 150 individual data points must be balanced with the costs of collecting the data. Non-parametric tests may be useful for analyzing smaller samples, but care must be taken to apply appropriate methods. So, before this study is repeated, consideration must be given to the issues of training, automation, and sample size.

Finally, the application of the AHP to other aggregation applications could be explored. Specifically, it might be successfully used to aggregate diverse cost, schedule, and performance measures for complex acquisition projects. Or, it could be used in its traditional role as a decision support tool for source selection, or any other complex decision which may be of import to the Air Force at the time of the research.

Despite the conclusions of this study, the AHP remains a very adaptable and robust process for logically considering many factors during decision making. The results of this study should not tarnish its reputation or diminish its application to tasks for which it was designed. Furthermore, the study did demonstrate a use of the AHP that may foster other innovative applications.

## Appendix A: Data Collection Briefing

# USA Using the AHP

Scoring Unit Self-Assessments (USA)
Using the
Analytic Hierarchy Process (AHP)

1

# Overview

- Purpose and Approach
- Analytic Hierarchy Process (AHP) Tutorial
- AHP-USA Scoring
- Quality Air Force (QAF) Process Review
- QAF-USA Scoring
  - Individual
  - Team consensus
- Feedback Questionnaire

2

## Non-attribution

- Your name will not be published or associated with this research.
- The unit self-assessment report that you will score, is provided with permission of the originating unit.
- The originating unit's name will not be published or associated with this research.

EXPRESS YOUR TRUE OPINION

3

## Purpose & Approach

- Purpose: To evaluate the feasibility and desirability of using the Analytic Hierarchy Process (AHP) to score unit self-assessments.
- Approach: Compare the AHP-based scoring method to the current Quality Air Force (QAF) scoring method.

| | |
|---|---|
| - Score | - Time |
| - Consistency | - Ease of use |

4

# AHP Tutorial
## Overview

- Background
- Example AHP application
  - Step 1: Create the hierarchy
  - Step 2: Evaluate the hierarchy
  - Step 3: Calculate the score for each alternative
- Unit self-assessment AHP application

5

# Background

- Measurement theory and decision-making process developed by Thomas L. Saaty in the early 1970s for military contingency planning.
- Many other applications and studies:

- Multicriterion decision making
- Mathematics
- Marketing
- Law
- Public Policy
- Economics
- Sports
- Medicine
- Transportation planning
- Conflict resolution

6

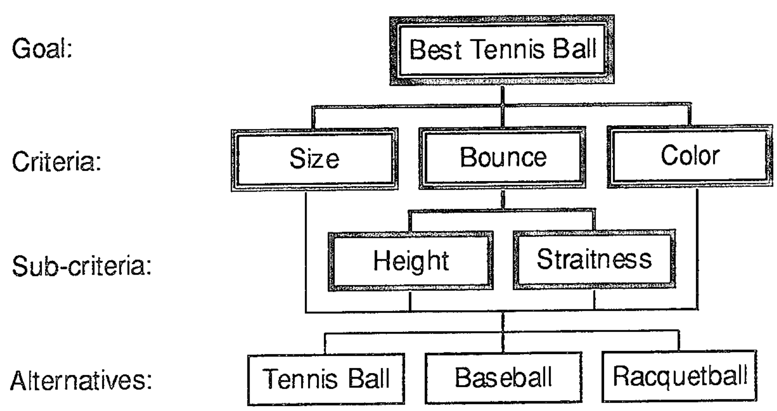
# AHP Example
# Selecting a Ball for Tennis

- Goal: From a set of three balls, choose the best one for playing tennis.
- Process:
  - Step 1: Identify evaluation criteria and alternatives to create a hierarchy.
  - Step 2: Evaluate the hierarchy to weight criteria and alternatives.
  - Step 3: Calculate a score for each alternative.

7

# Step 1: Create the Hierarchy

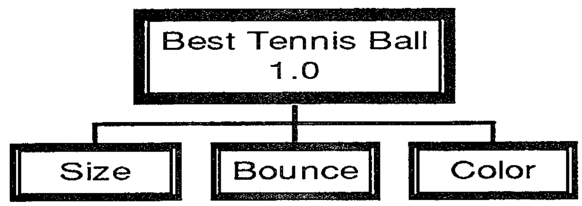
Selecting a Ball for Playing Tennis




8

# Step 2: Evaluate the Hierarchy

Evaluate the Criteria
with respect to the Goal



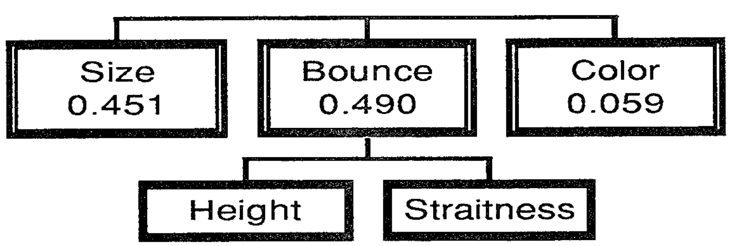

1=Equal 3=Moderate 5=Strong 7=Very Strong 9=Extreme

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Bounce |
| Size | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Color |
| Bounce | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Color |

9

# Step 2: Evaluating the Hierarchy

Evaluate Sub-criteria
with respect to the Criteria



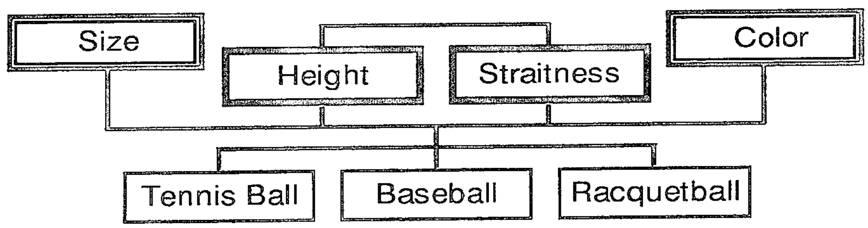

1=Equal 3=Moderate 5=Strong 7=Very Strong 9=Extreme

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Height | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Straight |

10

# Step 2: Evaluating the Hierarchy

Evaluate the Alternatives
with respect to the "Leaves"
(Criteria & Sub-criteria without Subordinates)




| | | | | |
|---|---|---|---|---|
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |

11

# My Comparisons

SIZE

| | | | | |
|---|---|---|---|---|
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |

HEIGHT

| | | | | |
|---|---|---|---|---|
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |

STRAIGHTNESS

| | | | | |
|---|---|---|---|---|
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |

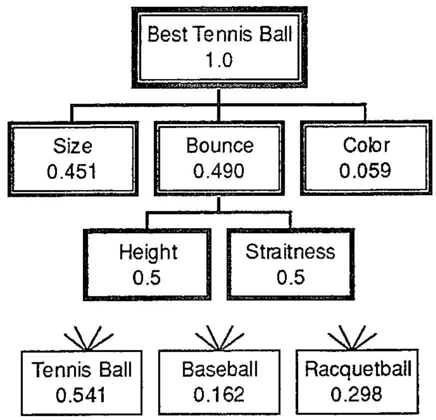
COLOR

| | | | | |
|---|---|---|---|---|
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquet |

12

# Step 3: Calculate Scores

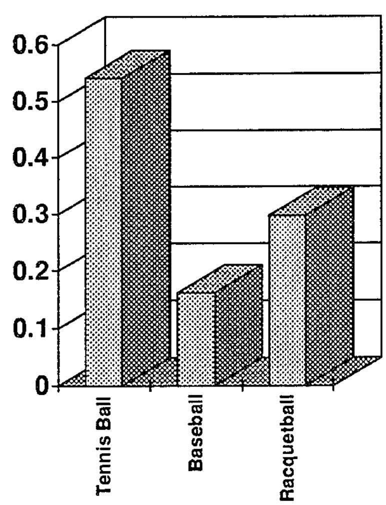
Selecting a Ball for Playing Tennis




13

# Final Scores

- Criteria weights are distributed to each alternative.
- Resulting scores are relative to the other competing alternatives.
- In this example, the tennis ball's score becomes the "ideal" or "perfect" value for comparison.




14

# AHP-USA Scoring

- Step 1: Create Hierarchy -- **done** -- created from QAF item 5.2 (Process Management: Product and Service Production and Delivery Processes).
- Step 2: Evaluate Hierarchy -- **your job** -- weight three areas, nine sub-areas, and then compare two alternatives with respect to the hierarchy's leaves:
  - Ideal = ideal organization based on QAF criteria.
  - Actual = case study unit self-assessment report.
- Step 3: Calculate Scores -- **my job**.

15

# Actual versus Ideal

- ACTUAL performance defined by:
- Unit self-assessment report:
  - Mission (KBF).
  - Performance (item 5.2).
- Your understanding of the USA report.

- IDEAL performance defined by:
- Quality Air Force documentation:
  - Categories, items, & areas.
  - Definition of 100% score.
- Your understanding of the QAF criteria & definition.

**Your Comparison Judgments**

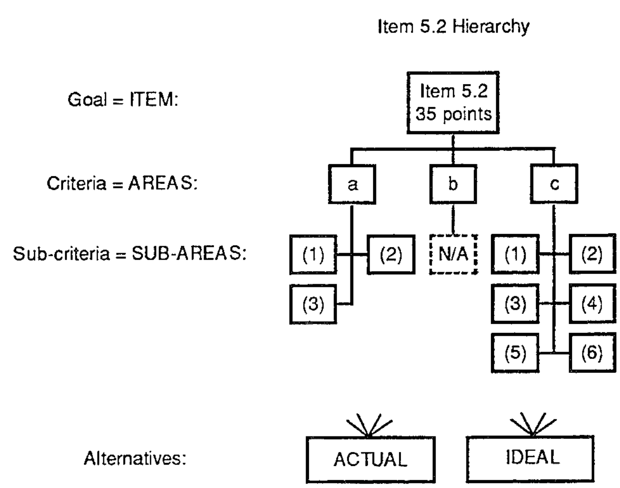
16

# IDEAL (100%) PERFORMANCE for Approach/Deployment Items

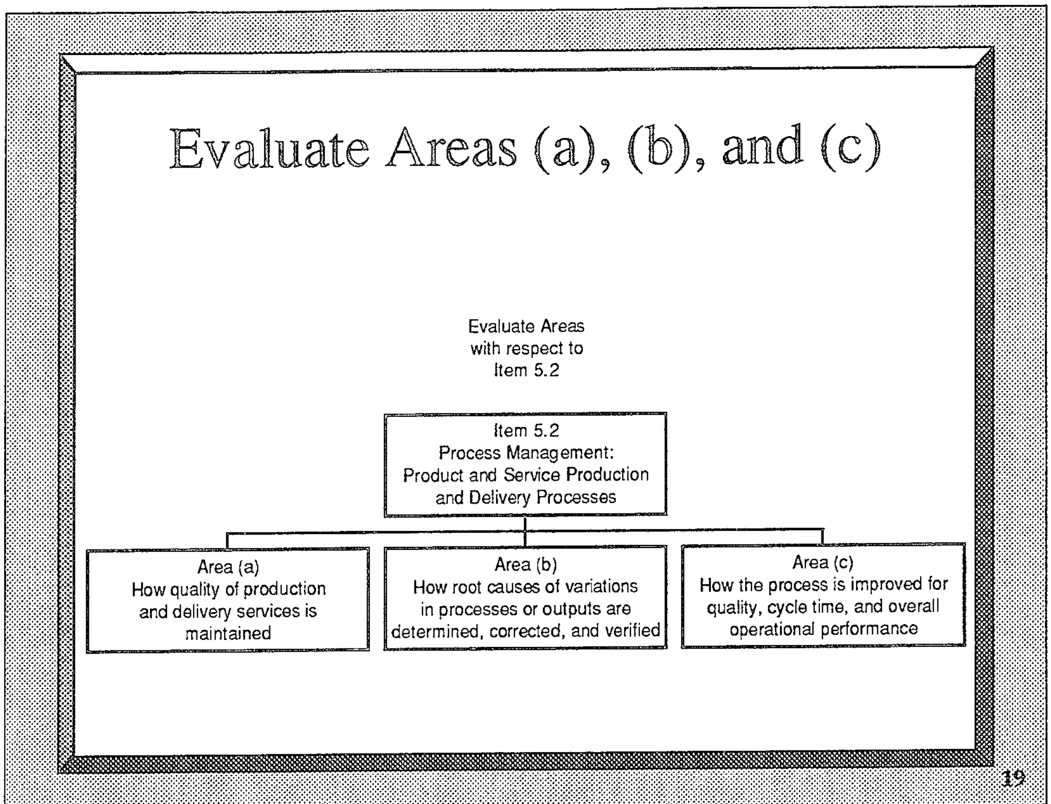
- **Sound, systematic approach** fully responsive to **all requirements** of the item.
- Approach is **fully deployed** without weaknesses or gaps in any areas.
- Very strong **refinement** and **integration** -- backed by excellent **analysis**.

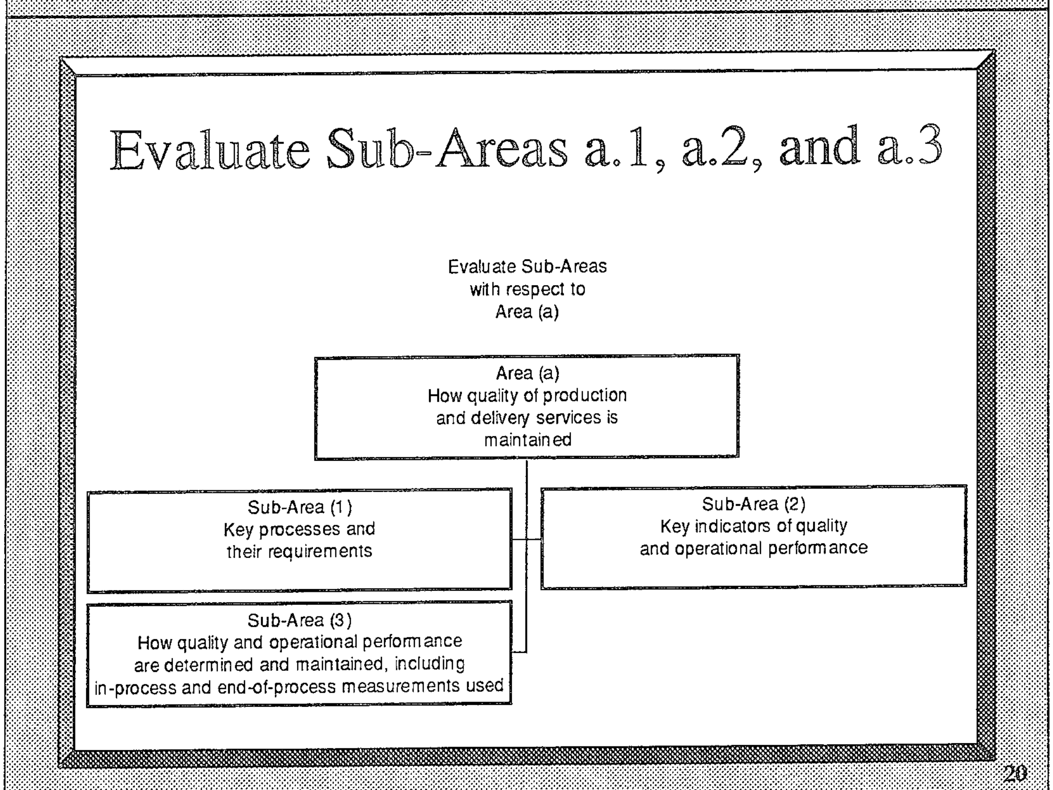
17

# Item 5.2 Hierarchy

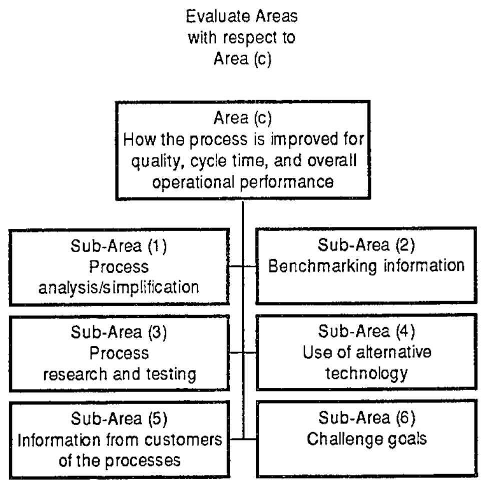



18


Evaluate Areas (a), (b), and (c)
Evaluate Areas with respect to Item 5.2
Item 5.2 Process Management: Product and Service Production and Delivery Processes
Area (a) How quality of production and delivery services is maintained
Area (b) How root causes of variations in processes or outputs are determined, corrected, and verified
Area (c) How the process is improved for quality, cycle time, and overall operational performance
19

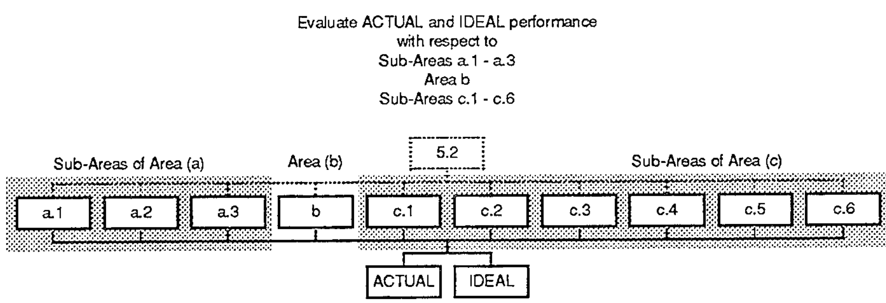

Evaluate Sub-Areas a.1, a.2, and a.3
Evaluate Sub-Areas with respect to Area (a)
Area (a) How quality of production and delivery services is maintained
Sub-Area (1) Key processes and their requirements
Sub-Area (2) Key indicators of quality and operational performance
Sub-Area (3) How quality and operational performance are determined and maintained, including in-process and end-of-process measurements used
20

# Evaluate Sub-Areas c.1 through c.6




21

# Evaluate Alternatives with respect to "Leaves"




22

# AHP-USA Scoring is Done !!

Please take a 10 minute break

23

# QAF Scoring Review

- Read/review the unit self-assessment package.
  - Key business factors.
  - Item 5.2.
- Identify strengths/areas for improvement.
- Individually score each item using the Baldrige scoring guidelines (percentage scale).
- Consensus score for each item where individual scores vary by more than 20%.

24

# Begin Individual Scoring

- Record your start/stop times for each section.
- 1 hour has been allocated for individual scoring; however, this is not a hard deadline.
- When you are done, take a break.

25

# QAF Scoring Guidelines

| | |
|---|---|
| ■ **No system** evident, **anecdotal** information. | **0%** |
| ■ **Beginning of a systematic approach** to addressing **primary** purposes of the item.<br>■ **Significant gaps** still exist in deployment.<br>■ **Early stages** of transition from reacting to **preventing** problems. | **10%** \| **30%** |
| ■ **Sound, systematic approach** responsive to the **primary** purposes of the item.<br>■ **Fact-based improvement** process in place.<br>■ **No major gaps** in deployment, though some areas may be in early stages.<br>■ **More emphasis** placed on problem **prevention** than reaction to problems. | **40%** \| **60%** |
| ■ Sound, systematic approach responsive to the **overall** purposes of the item.<br>■ Fact-based improvement process is a **key management tool; evidence of refinement** as a result of improvement cycles and analysis.<br>■ **Well deployed** with no significant gaps, although refinement, deployment, and integration may vary among work units. | **70%** \| **90%** |
| ■ Sound, systematic approach **fully responsive** to **all requirements** of the item.<br>■ Approach is **fully deployed** without weaknesses or gaps in any areas.<br>■ Very strong **refinement** and **integration** -- backed by excellent **analysis**. | **100%** |

26

# Begin Team Consensus Scoring

- If your team does not require consensus, because individual scores are within a 20% range, then I will provide instructions on completing the feedback survey after starting the other teams.
- Use the flip charts to display all strengths and areas for improvement.
- Record the start/stop times for your session.

27

# Feedback Questionnaire

- Turn to page 12 of the data collection package.
- Answer the questions and comments:
  - Paired comparisons like those used to evaluate the hierarchy.
  - Comments.

28

## Completion Checklist

- Please turn to the last page (pg 14) of the data collection package.
- Verify that all items have been completed.
- **Team leaders** need to verify that at least one team consensus scoring sheet is completely filled out.

29

## THANK YOU !

In appreciation for your time and effort,
I am offering free soda/beer at the
Flywright for lunch!

30

## Appendix B: Data Collection Package

# Unit Self-Assessment (USA)

# Using the

# Analytical Hierarchy Process (AHP)

# Data Collection Package

## Overview

- Purpose and approach of this study.
- Analytic hierarchy process (AHP) tutorial and example.
- AHP scoring of USA report.
- Quality Air Force (QAF) process review.
- QAF scoring of USA report.
- Process evaluation questionnaire.

## Non-attribution

- Your name will not be published or associated with this research.
- The unit self-assessment report that you will score in this study, is provided with permission of the originating unit.
- The originating unit's name will not be published or associated with this research.

PLEASE FEEL FREE TO EXPRESS YOUR TRUE OPINION

## AHP Tutorial and Example (Picking up at Step 2)

## Step 2: Evaluate the Hierarchy

### 2.1 Evaluate the criteria with respect to the goal.

Which criteria (size, bounce, and color) is more important with respect to the goal?

**(Goal):** **Best tennis ball.**

| | | | | |
|---|---|---|---|---|
| Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Bounce |
| Size | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Color |
| Bounce | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Color |

### 2.2 Evaluate the sub-criteria with respect to the criteria.

Which sub-criteria (height and straightness) is more important with respect to the criteria?

**(Criteria):** **Bounce.**

| | | | | |
|---|---|---|---|---|
| Height | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Straightness |

### 2.3 Evaluate the alternatives with respect to the "leaves."

Which is the preferred alternative with respect to SIZE?

| | | | | |
|---|---|---|---|---|
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |

Which is the preferred alternative with respect to HEIGHT?

| | | | | |
|---|---|---|---|---|
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |

Which is the preferred alternative with respect to STRAIGHTNESS?

| | | | | |
|---|---|---|---|---|
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |

Which is the preferred alternative with respect to COLOR?

| | | | | |
|---|---|---|---|---|
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Baseball |
| Tennis Ball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |
| Baseball | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Racquetball |

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

# AHP-USA Scoring

## Team Data

Team number: ____ You have been designated a **team leader** if this box is marked: ☐

## Personal Data

Name: ______________________________

Work Phone: ___________________ E-mail: ______________________________

Have you had unit self-assessment (USA) training?

- ☐ Yes
- ☐ No

If you answered YES to question 1, how long ago were you trained?

- ☐ Less than 3 months ago
- ☐ 3 - 6 months
- ☐ 6 - 12 months
- ☐ Over 12 months

What experience do you have with USA (check all that apply):

- ☐ No experience, other than training.
- ☐ Participated in data collection/preparation of self-assessment reports for your unit.
- ☐ Scored USA reports.
- ☐ Visited unit(s) as a USA consultant (for feedback to the unit).
- ☐ Visited unit(s) as a USA examiner (for award application site visit).
- ☐ Other: ____________________________________________________

## Overview

Step 1: Create the Hierarchy -- Already done, based on the 1993 Quality Air Force criteria.

Step 2: Evaluate the Hierarchy

> Step 2.1: Evaluate the criteria (areas a, b, and c) with respect to the goal (item 5.2).
> Step 2.2: Evaluate the sub-criteria (sub-areas a.1, a.2, a.3) with respect to the criteria (area a).
> Step 2.3: Evaluate the sub-criteria (sub-areas c.1 through c.6) with respect to the criteria (area c).
> Step 2.4: Evaluate the alternatives (actual and ideal) with respect to the “leaves” (a.1...a.3, b, c.1...c.6).

Step 3: Calculate Scores for each Alternative -- Will be done as part of data analysis.

## Instructions

1. Fill out the personal data above.
2. Record start/stop times in the spaces provided as you begin and finish scoring sections.
3. If you prefer word-based comparisons over numerical comparisons, then refer to the scale shown at the bottom of each page (see below).
4. Don’t hesitate to ask questions.

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

## Step 1: Create the Hierarchy

Already done, based on the 1993 Quality Air Force criteria for Item 5.2.

## Step 2: Evaluate the Hierarchy

### 2.1 Evaluate the criteria (areas a, b, and c) with respect to the goal (item 5.2).

PLEASE RECORD THE TIME WHEN YOU BEGIN: __________

Which area (a, b, and c) is more important with respect to item 5.2?

**(5.2): Process Management: Product and Service Production and Delivery Processes**

| | | | | |
|---|---|---|---|---|
| **(a):** How quality of production and delivery services is maintained | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(b):** How root causes of variations in processes or outputs are determined, corrected and verified |
| **(a):** How quality of production and delivery services is maintained | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(c):** How the process is improved for quality, cycle time, and overall operational performance |
| **(b):** How root causes of variations in processes or outputs are determined, corrected and verified | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(c):** How the process is improved for quality, cycle time, and overall operational performance |

PLEASE RECORD THE TIME WHEN YOU ARE DONE: __________

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

## Step 2: Evaluate the Hierarchy...

**2.2 Evaluate the sub-criteria (sub-areas a.1, a.2, and a.3) with respect to the criteria (area a).**

PLEASE RECORD THE TIME WHEN YOU BEGIN: _________

Which **sub-area** (1, 2, and 3) is more important with respect to **area** (a)?

**(a): How quality of production and delivery services is maintained.**

| | | | | |
|---|---|---|---|---|
| **(1):** Key processes and their requirements | **9** 8 **7** 6 **5** 4 **3** 2 | **1** | 2 **3** 4 **5** 6 **7** 8 **9** | **(2):** Key indicators of quality and operational performance |
| **(1):** Key processes and their requirements | **9** 8 **7** 6 **5** 4 **3** 2 | **1** | 2 **3** 4 **5** 6 **7** 8 **9** | **(3):** How quality and operational performance are determined and maintained, including the in-process and end-of-process measurements used |
| **(2):** Key indicators of quality and operational performance | **9** 8 **7** 6 **5** 4 **3** 2 | **1** | 2 **3** 4 **5** 6 **7** 8 **9** | **(3):** How quality and operational performance are determined and maintained, including the in-process and end-of-process measurements used |

PLEASE RECORD THE TIME WHEN YOU ARE DONE: _________

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

## Step 2: Evaluate the Hierarchy...

**2.3 Evaluate the sub-criteria (sub-areas c.1 through c.6) with respect to the criteria (area c).**

PLEASE RECORD THE TIME WHEN YOU BEGIN: __________

Which sub-area (1 through 6) is more important with respect to area (c)?

**(c): How the process is improved for quality, cycle time, and overall operational performance.**

| | | | | |
|---|---|---|---|---|
| **(1)**: Process analysis/simplification | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(2)**: Benchmarking information |
| **(1)**: Process analysis/simplification | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(3)**: Process research and testing |
| **(1)**: Process analysis/simplification | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(4)**: Use of alternative technology |
| **(1)**: Process analysis/simplification | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(5)**: Information from customers of the process |
| **(1)**: Process analysis/simplification | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(6)**: Challenge goals |
| **(2)**: Benchmarking information | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(3)**: Process research and testing |
| **(2)**: Benchmarking information | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(4)**: Use of alternative technology |
| **(2)**: Benchmarking information | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(5)**: Information from customers of the process |
| **(2)**: Benchmarking information | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(6)**: Challenge goals |
| **(3)**: Process research and testing | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(4)**: Use of alternative technology |
| **(3)**: Process research and testing | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(5)**: Information from customers of the process |
| **(3)**: Process research and testing | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(6)**: Challenge goals |
| **(4)**: Use of alternative technology | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(5)**: Information from customers of the process |
| **(4)**: Use of alternative technology | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(6)**: Challenge goals |
| **(5)**: Information from customers of the process | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | **(6)**: Challenge goals |

PLEASE RECORD THE TIME WHEN YOU ARE DONE: __________

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

## Step 2: Evaluate the Hierarchy...

**2.4 Evaluate the alternatives (actual and ideal) with respect to the "leaves" (a.1...a.3, b, c.1...c.6).**

Note: You should **not** mark the shaded side because the ACTUAL should never be preferred over the IDEAL. In other words, the **best** a unit can do is attain **equal preference [1]** when compared to the ideal.

**Reminder of the definitions of higher-level criteria:**
(5.2): Process Management: Product and Service Production and Delivery Processes
(a): How quality of production and delivery services is maintained.

PLEASE RECORD THE TIME WHEN YOU BEGIN: __________

Which **alternative** is preferred with respect to **sub-area a.1**?

**(a.1): Key processes and their requirements.**

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which **alternative** is preferred with respect to **sub-area a.2**?

**(a.2): Key indicators of quality and operational performance.**

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which **alternative** is preferred with respect to **sub-area a.3**?

**(a.3): How quality and operational performance are determined and maintained, including in-process and end-of-process measurements used.**

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which **alternative** is preferred with respect to **area b**?

**(b): How root causes of variations in processes or outputs are determined, corrected, and verified.**

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

PLEASE CONTINUE EVALUATIONS ON NEXT PAGE

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

2.4 Evaluating the alternatives with respect to the leaves, continued...

Reminder of the definitions of higher-level criteria:
(5.2): Process Management: Product and Service Production and Delivery Processes
(c): How the process is improved for quality, cycle time, and overall operational performance.

Which alternative is preferred with respect to sub-area c.1?
(c.1): Process analysis/simplification.

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which alternative is preferred with respect to area c.2?
(c.2): Benchmarking information.

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which alternative is preferred with respect to sub-area c.3?
(c.3): Process research and testing.

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which alternative is preferred with respect to sub-area c.4?
(c.4): Use of alternative technology.

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which alternative is preferred with respect to sub-area c.5?
(c.5): Information from customers of the processes.

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

Which alternative is preferred with respect to sub-area c.6?
(c.6): Challenge goals.

| | | | | |
|---|---|---|---|---|
| Actual | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | Ideal |

PLEASE RECORD THE TIME WHEN YOU ARE DONE: __________

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

# QAF-USA Scoring

## Overview

Step 1: Individual scoring.

Step 1.1: Read the USA report and identify strengths (+) and areas for improvement (-).
Step 1.2: Use the QAF percentage scale to assign a score for the item.

Step 2: Consensus scoring.

Step 2.1: Determine whether your individual scores vary by more than 20%.
Step 2.2: If so, then work with your team members to review strengths and areas for improvement in order to reach consensus.

Step 3: Record your individual and consensus scores in the spaces provided.

## Instructions

1. Review the steps described above, if you haven't already done so.
2. Record start/stop times in the spaces provided as you begin and finish individual and consensus scoring.
3. Don't hesitate to ask questions.

## QAF-USA Individual Scoring Worksheet

PLEASE RECORD THE TIME WHEN YOU BEGIN: _________

5.2 Process Management: Product and Service Production and Delivery Processes (35 points)

| + / ++ | Area to Address | (+) STRENGTHS |
|---|---|---|
| | | |
| **- / --** | **Area to Address** | **(-) AREAS FOR IMPROVEMENT** |
| | | |

INDIVIDUAL PERCENT SCORE:

PLEASE RECORD THE TIME WHEN YOU ARE DONE: _________

## QAF-USA Team Consensus Scoring Worksheet

PLEASE RECORD THE TIME WHEN YOUR **TEAM** BEGINS: __________

5.2 Process Management: Product and Service Production and Delivery Processes (35 points)

| + / ++ | Area to Address | (+) STRENGTHS |
|---|---|---|
| | | |
| **- / --** | **Area to Address** | **(-) AREAS FOR IMPROVEMENT** |
| | | |

| INDIVIDUAL PERCENT SCORES | **BEFORE** and | **AFTER** | CONSENSUS |
|---|---|---|---|
| Name 1:__________________ | | | Average the individual scores after consensus to find your: |
| Name 2:__________________ | | | |
| Name 3:__________________ | | | TEAM CONSENSUS |
| Name 4:__________________ | | | |
| Name 5:__________________ | | | |

PLEASE RECORD THE TIME WHEN YOUR **TEAM** IS DONE: __________

# Feedback Questionnaire

## Understandability

Which method is easier to understand?

| AHP<br>-Hierarchy<br>-Paired comparisons | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF<br>- Categories, items, areas<br>- Points and percentages |
|---|---|---|---|---|

Which method is more intuitive?

| AHP<br>-Hierarchy<br>-Paired comparisons | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF<br>- Categories, items, areas<br>- Points and percentages |
|---|---|---|---|---|

Which method is simpler?

| AHP<br>-Hierarchy<br>-Paired comparisons | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF<br>- Categories, items, areas<br>- Points and percentages |
|---|---|---|---|---|

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

**Usability**

For which method was it easier to determine a score for the unit?

| | | | | |
|---|---|---|---|---|
| AHP<br>- Paired comparisons of alternatives with respect to the 10 "leaves" | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF<br>- Individual scoring<br>- Consensus scoring |

Which method was easier to use overall in this scoring exercise?

| | | | | |
|---|---|---|---|---|
| AHP<br>- Paired comparisons for the whole hierarchy (criteria, sub-criteria, alternatives) | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF<br>- Individual scoring<br>- Consensus scoring |

Which method was a better tool to identify strengths and areas for improvement for the unit?

| | | | | |
|---|---|---|---|---|
| AHP<br>- Paired comparisons of alternatives with respect to the 10 "leaves" | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF<br>- Plus/Minus "scoring" while reading report |

**Believability**

If an "accurate score" is defined as "a score which truly represents a unit's performance relative to desired/planned performance," which method do you believe would produce the most accurate scores?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

If a "consistent score" is defined as "a score which varies little between evaluators," which method do you believe would produce the most consistent scores?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

Which method would you trust more to score unit self-assessments?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

## Applicability

Which method would you prefer to use for future unit self-assessments?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

Which method would you prefer to accurately show a trend in USA scores from year to year?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

Which method would you prefer if USA were to be done by novice practitioners?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

Which method would you prefer if USA were to be done by experienced practitioners?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

Which method is best suited to a variety of USA practitioners (novice to experienced)?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

Which method do you prefer overall?

| | | | | |
|---|---|---|---|---|
| AHP | 9 8 7 6 5 4 3 2 | 1 | 2 3 4 5 6 7 8 9 | QAF |

## Comments

Please continue on the back, if you need more room.

Degrees of Preference: 1 = Equal 3 = Moderate 5 = Strong 7 = Very Strong 9 = Extreme (2, 4, 6, and 8 are in-between values)

## Completion Checklist

PLEASE VERIFY THE FOLLOWING ITEMS ARE COMPLETE:

- ☐ Personal data is complete (page 136).
- ☐ AHP-USA paired comparisons are done and legible (pages 137-142).
- ☐ QAF-USA individual score is recorded (page 143).
- ☐ At least one QAF-USA team consensus score sheet is complete for your team (page 145).
- ☐ Feedback questionnaire is complete (pages 146-148).

PLEASE LEAVE THIS PACKAGE AT YOUR TABLE WHEN YOU ARE DONE.

**THANK YOU FOR YOUR TIME AND EFFORT !!**

## Appendix C: Results from Feedback Questions

TABLE 32

RESPONSES TO UNDERSTANDABILITY QUESTION 1

Which method is easier to understand?

| Eval # | AHP | | | | | | | EQUAL | | | QAF | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | | | T | | | | |
| 2 | | | | | | | | | | | T | | | | | | |
| 3 | | | T | | | | | | | | | | | | | | |
| 4 | | | | | | | U | | | | | | | | | | |
| 5 | | | | | | | | | | T | | | | | | | |
| 6 | | | T | | | | | | | | | | | | | | |
| 7 | | | | | U | | | | | | | | | | | | |
| 8 | | | | | U | | | | | | | | | | | | |
| 9 | | | | | | | T | | | | | | | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | U | | | | | | | | | | | | |
| Total | 0 | 0 | 2 | 0 | 3 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |

TABLE 33

RESPONSES TO UNDERSTANDABILITY QUESTION 2

Which method is more intuitive?

| Eval # | AHP | | | | | | | EQUAL | | | QAF | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | T | | | | | | | | | | |
| 2 | | | | | | | | | | | T | | | | | | |
| 3 | | | T | | | | | | | | | | | | | | |
| 4 | | | | | | | U | | | | | | | | | | |
| 5 | | | | | | | | | | | T | | | | | | |
| 6 | | | T | | | | | | | | | | | | | | |
| 7 | | | | | | | | | | U | | | | | | | |
| 8 | | | | | U | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | | T | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | | | U | | | | | | | | | | |
| Total | 0 | 0 | 2 | 0 | 1 | 0 | 3 | 0 | 1 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |

## TABLE 34

## RESPONSES TO UNDERSTANDABILITY QUESTION 3

Which method is simpler?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | QAF | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | T | | | | | | | | | |
| 2 | | | | | | | T | | | | | | | | | | |
| 3 | | | T | | | | | | | | | | | | | | |
| 4 | | | | | | | U | | | | | | | | | | |
| 5 | | | | | | | | | | T | | | | | | | |
| 6 | T | | | | | | | | | | | | | | | | |
| 7 | | | | | U | | | | | | | | | | | | |
| 8 | | | U | | | | | | | | | | | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | U | | | | | | | | | | | | |
| Total | 1 | 0 | 2 | 0 | 3 | 0 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## TABLE 35

## RESPONSES TO USABILITY QUESTION 1

For which method was it easier to determine a score for the unit?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | QAF | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | | | T | | | | |
| 2 | | | | | | | | | | T* | | | | | | | |
| 3 | | | T | | | | | | | | | | | | | | |
| 4 | | | | | | | U | | | | | | | | | | |
| 5 | | | | | | | | | | | T | | | | | | |
| 6 | | | | | | | | | | | | | T | | | | |
| 7 | | | | | U | | | | | | | | | | | | |
| 8 | | | | | | | | | U | | | | | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | | | | | T | | | | | | |
| 11 | | | | | | | U | | | | | | | | | | |
| Total | 0 | 0 | 1 | 0 | 2 | 0 | 2 | 0 | 1 | 1 | 2 | 0 | 2 | 0 | 0 | 0 | 0 |

* Evaluator two marked the space between values of 2 and 3, therefore a value of 2.5 was used to calculate the geometric mean and a single “2” vote was used for the histogram.

## TABLE 36

## RESPONSES TO USABILITY QUESTION 2

Which method was easier to use overall in this scoring exercise?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | T | | | | | | |
| 2 | | | | | | | | | | | | | T | | | | |
| 3 | | | T | | | | | | | | | | | | | | |
| 4 | | | | | U | | | | | | | | | | | | |
| 5 | | | | | | | | | | | T | | | | | | |
| 6 | | | | | | | | | | | | | T | | | | |
| 7 | | | | | U | | | | | | | | | | | | |
| 8 | | | | | U | | | | | | | | | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | | U | | | | | | | | | | | |
| Total | 0 | 0 | 1 | 0 | 4 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 |

## TABLE 37

## RESPONSES TO USABILITY QUESTION 3

Which method was a better tool to identify strengths and areas for improvement for the unit?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | | | | | | | T |
| 2 | | | | | | | | | | | | | | | T | | |
| 3 | | | | | | | | | T | | | | | | | | |
| 4 | | | | | | | | | U | | | | | | | | |
| 5 | | | | | | | | | | | | | | | | | T |
| 6 | | | | | | | T | | | | | | | | | | |
| 7 | | | | | | | | | | | U | | | | | | |
| 8 | | | | | | | | | | | | | | | U | | |
| 9 | | | | | | | | | | | | | | | T | | |
| 10 | | | | | | | | | | | T | | | | | | |
| 11 | | | | | | | | | | | | U | | | | | |
| Total | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 2 |

## TABLE 38

## RESPONSES TO BELIEVABILITY QUESTION 1

Which method do you believe will produce the most "accurate" scores?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | T | | | | | | | |
| 2 | | | | | | | T | | | | | | | | | | |
| 3 | | | | | | | T | | | | | | | | | | |
| 4 | | | | | | | | U | | | | | | | | | |
| 5 | | | | | | | | | | | | | T | | | | |
| 6 | | | | | T | | | | | | | | | | | | |
| 7 | | | | | | | | | | U | | | | | | | |
| 8 | | | | | | | | | | | | | U | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | T | | | | | | | | | | |
| 11 | | | | | | U | | | | | | | | | | | |
| Total | 0 | 0 | 0 | 0 | 2 | 1 | 3 | 1 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |

## TABLE 39

## RESPONSES TO BELIEVABILITY QUESTION 2

Which method do you believe will produce the most "consistent" scores?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | T | | | | | | | | | |
| 2 | | | | | | | T | | | | | | | | | | |
| 3 | | | | | | | T | | | | | | | | | | |
| 4 | | | U | | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | T | | | | | | |
| 6 | | | | | T | | | | | | | | | | | | |
| 7 | | | | | U | | | | | | | | | | | | |
| 8 | | | | | | | | | | | | | U | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | | | | | U | | | | | | | | |
| Total | 0 | 0 | 1 | 0 | 3 | 0 | 2 | 1 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |

## TABLE 40

## RESPONSES TO BELIEVABILITY QUESTION 3

Which method would you trust more to score unit self-assessments?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | | | | | T | | |
| 2 | | | | | | | | | | | T | | | | | | |
| 3 | | | | | | | T | | | | | | | | | | |
| 4 | | | | | | | U | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | | | | T |
| 6 | | | | | T | | | | | | | | | | | | |
| 7 | | | | | | | | | U | | | | | | | | |
| 8 | | | | | | | | | | | | | U | | | | |
| 9 | | | | | | | T | | | | | | | | | | |
| 10* | | | | | | | | | | | | | | | | | |
| 11 | | | | | | | U | | | | | | | | | | |
| Total | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |

* Evaluator ten abstained from making this judgment.

## TABLE 41

## RESPONSES TO APPLICABILITY QUESTION 1

Which method would you prefer to use for future unit self-assessments?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | | | | T | | | |
| 2 | | | | | | | | | T | | | | | | | | |
| 3 | | | | T | | | | | | | | | | | | | |
| 4 | | | | | U | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | T | | | | |
| 6 | | | | | T | | | | | | | | | | | | |
| 7 | | | | | | U | | | | | | | | | | | |
| 8 | | | | | | | | | | | | | U | | | | |
| 9 | | | | | | | T | | | | | | | | | | |
| 10 | | | | | | | | | | | T | | | | | | |
| 11 | | | | | | | U | | | | | | | | | | |
| Total | 0 | 0 | 0 | 1 | 2 | 1 | 2 | 0 | 1 | 0 | 1 | 0 | 2 | 1 | 0 | 0 | 0 |

## TABLE 42

## RESPONSES TO APPLICABILITY QUESTION 2

Which method would you prefer to accurately show a trend from year to year?

| Eval # | AHP | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | T | | | | | | | | | |
| 2 | | | | | | | | | T | | | | | | | | |
| 3 | | | | | T | | | | | | | | | | | | |
| 4 | | | | | | | U | | | | | | | | | | |
| 5 | | | | | | | | | | | T | | | | | | |
| 6 | | | | | | | | | T | | | | | | | | |
| 7 | | | | | | | | | | | U | | | | | | |
| 8 | | | | | | | | | U | | | | | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | | | U | | | | | | | | | | |
| Total | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 1 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |

## TABLE 43

## RESPONSES TO APPLICABILITY QUESTION 3

Which method would you prefer if USA were to be done by novice practitioners?

| Eval # | AHP | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | T | | | | | | | | | |
| 2 | | | | | | | | | T | | | | | | | | |
| 3 | | | | T | | | | | | | | | | | | | |
| 4 | | | U | | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | T | | | | | | |
| 6 | | | | | T | | | | | | | | | | | | |
| 7 | | | U | | | | | | | | | | | | | | |
| 8 | | | | | | | U | | | | | | | | | | |
| 9 | | | T | | | | | | | | | | | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | U | | | | | | | | | | | | |
| Total | 0 | 0 | 3 | 1 | 2 | 0 | 1 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

## TABLE 44

## RESPONSES TO APPLICABILITY QUESTION 4

Which method would you prefer if USA were to be done by experienced practitioners?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | | | | | | T | |
| 2 | | | | | | | | | T | | | | | | | | |
| 3 | | | | | | T | | | | | | | | | | | |
| 4 | | | | | | | | | | | | | | | | | U |
| 5 | | | | | | | | | | | | | | | | | T |
| 6 | | | | | T | | | | | | | | | | | | |
| 7 | | | | | | | | | | | | | | | U | | |
| 8 | | | | | | | | | | | | | | | U | | |
| 9 | | | | | | | T | | | | | | | | | | |
| 10 | | | | | | | | | | | T | | | | | | |
| 11 | | | | | | | | | U | | | | | | | | |
| Total | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 2 |

## TABLE 45

## RESPONSES TO APPLICABILITY QUESTION 5

Which method is best suited to a variety of USA practitioners (novice to experienced)?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | T | | | | | | | |
| 2 | | | | | | | | | T | | | | | | | | |
| 3 | | | | | T | | | | | | | | | | | | |
| 4 | | | | | | | | | U | | | | | | | | |
| 5 | | | | | | | | | | | | | | T | | | |
| 6 | | | | | T | | | | | | | | | | | | |
| 7 | | | | | | U | | | | | | | | | | | |
| 8 | | | | | U | | | | | | | | | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | | | T | | | | | | | | |
| 11 | | | | | | | U | | | | | | | | | | |
| Total | 0 | 0 | 0 | 0 | 4 | 1 | 1 | 0 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

## TABLE 46

## RESPONSES TO APPLICABILITY QUESTION 6

Which method do you prefer overall?

| Eval # | AHP | | | | | | | | EQUAL | | | | | | | | QAF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | | | | | | | | | | | T | | | | | | |
| 2 | | | | | | | | | | | | | T | | | | |
| 3 | | | | | T | | | | | | | | | | | | |
| 4 | | | | | | | U | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | | | | T |
| 6 | | | T | | | | | | | | | | | | | | |
| 7 | | | | | U | | | | | | | | | | | | |
| 8 | | | | | | | | | U | | | | | | | | |
| 9 | | | | | T | | | | | | | | | | | | |
| 10 | | | | | | | | | | | T | | | | | | |
| 11 | | | | | | | U | | | | | | | | | | |
| Total | 0 | 0 | 1 | 0 | 3 | 0 | 2 | 0 | 1 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 1 |

## Bibliography


Aeronautical Systems Division (ASD). Total Quality in Aeronautical Systems Division. Wright-Patterson AFB: HQ ASD, 1990.

Aeronautical Systems Division (ASD). Improving the Winning Edge. Wright-Patterson AFB: HQ ASD, 1991.

Aeronautical Systems Division (ASD). ASD Pamphlet 700-8. Wright-Patterson AFB: HQ ASD, 30 April 1992.

Aguayo, Rafael. Dr. Deming: The American Who Taught the Japanese about Quality. New York: Carol Publishing Group, 1990.

Air Force Quality Center. The Quality Approach. Maxwell AFB: USAF, 1993.

Air Force Quality Institute (AFQI). Quality Air Force Criteria. Montgomery: GPO, 1993.

Apostolou, Barbara and John M. Hassell, "An empirical examination of the sensitivity of the analytic hierarchy process to departures from recommended consistency ratios," Mathematical and Computer Modelling 17: 163-170 (1993).

Brinkerhoff, Robert O. and Dennis E. Dressler. Productivity Measurement: A Guide for Managers and Evaluators. Newbury Park: SAGE Publications, Inc., 1990.

Brown, Mark Graham. Baldrige Award Winning Quality: How to Interpret the Malcolm Baldrige Award Criteria (Third Edition). White Plains: Quality Resources, 1993.

Cooper, Donald R. and C. William Emory. Business Research Methods (Fifth Edition). Chicago: Richard D. Irwin, Inc., 1995.

Downs, George W. and Patrick D. Larkey. The Search for Government Efficiency From Hubris to Helplessness. Philadelphia: Temple University Press, 1986.

Epstein, Michael K. and John C. Henderson. "Data Envelopment Analysis for Managerial Control and Diagnosis," Decision Sciences 20: 90-119 (Winter 1989).

Federal Quality Institute. The President's Quality Award Application. Washington: GPO, June 1995.

Federal Quality Institute. Self-Assessment Guide for Organizational Performance and Customer Satisfaction. Washington: GPO, December 1993.

Forman, Ernest H., "Facts and Fictions about the Analytic Hierarchy Process," Mathematical and Computer Modelling 17: 19-26 (1993).

Guilfoos, Stephen J. "Measuring Transfer Effectiveness or Why Don Quixote Tilts at Windmills." Presented at the Technology Transfer Society 1994 Annual Conference, 22-24 June 1994, Updated 1 July 1994.

Nahmias, Steven. Production and Operations Analysis (Second Edition). Burr Ridge: Richard D. Irwin, Inc., 1993.

Nicholas, John M. Managing Business and Engineering Projects. Englewood Cliffs: Prentice Hall, 1990.

Provost, Lloyd and Susan Leddick, "How to Take Multiple Measures to get a Complete Picture of Organizational Performance," National Productivity Review: 477-490 (Autumn 1993).

Saaty, Thomas L. The Analytic Hierarchy Process. New York: McGraw-Hill, Inc., 1980.

Saaty, Thomas L. Decision Making for Leaders. Belmont: Wadsworth, Inc., 1982.

Saaty, Thomas L., "Axiomatic Foundation of the Analytic Hierarchy Process," Management Science, 32: 841-855 (1986).

Saaty, Thomas L., "The Analytic Hierarchy Process -- What it is and How it is Used," Mathematical Modelling, 9: 161-176 (1987).

Saaty, Thomas L., "What is Relative Measurement? The Ratio Scale Phantom," Mathematical and Computer Modelling 17: 1-12 (1993).

Saaty, Thomas L., "Highlights and Critical Points in the Theory and Application of the Analytic Hierarchy Process," European Journal of Operational Research, 74: 426-427 (1994).

Saaty, Thomas L. and Luis G. Vargas. The Logic of Priorities. Boston: Kluwer-Nijhoff Publishing, 1982.

Sink, D. Scott. Productivity Management: Planning, Measurement and Evaluation, Control and Improvement. New York: John Wiley & Sons, 1985.

Vargas, Luis G., "An Overview of the Analytic Hierarchy Process and its Applications," European Journal of Operational Research 48: 2-8 (1990)

Wedley, William C., "Consistency Prediction for Incomplete AHP Matrices," Mathematical and Computer Modelling 17: 151-161 (1993).

Wipper, Laura R., "Oregon Department of Transportation Steers Improvement with Performance Measurement," National Productivity Review 13: 359-367 (Summer 1994)

Young, Scott T., "Multiple Productivity Measurement Approaches for Management," Health Care Management Review 17(2): 51-58 (1992).

# Vita

Captain Bret L. Indermill was born on 16 May 1963, in Boulder Colorado to his parents: Roy C., and Kathryn K. Indermill. He graduated from Boulder High School in 1981 and matriculated at the University of Colorado in the fall of that year. In 1985, he received his Bachelor of Science Degree in Aerospace Engineering and was commissioned a Second Lieutenant in the United States Air Force. His first assignment took him to the 544th Strategic Intelligence Wing, Trajectory Division, located at Offutt Air Force Base (AFB) in Omaha Nebraska. There, he spent five and one half years working as a Trajectory Engineer, Single Integrated Operational Plan (SIOP) Support Engineer, and Chief of Configuration Management, supporting all US Intercontinental Ballistic Missile (ICBM) weapon systems. Then in 1991, Captain Indermill was reassigned as a project manager to a classified system program office located at Los Angeles AFB in El Segundo, California. There, he spent three years managing computer hardware and software acquisition and development contracts. Finally, in 1994 Captain Indermill entered the School of Logistics and Acquisition Management at the Air Force Institute of Technology (AFIT) to pursue a degree in Systems Management.

Permanent Address:
787 18th Street
Boulder CO 80302

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. AGENCY USE ONLY *(Leave blank)* | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | September 1995 | Master's Thesis |

**4. TITLE AND SUBTITLE**

AGGREGATING ORGANIZATIONAL PERFORMANCE METRICS USING THE ANALYTIC HIERARCHY PROCESS

**5. FUNDING NUMBERS**

**6. AUTHOR(S)**

Bret L. Indermill, Captain, USAF

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Air Force Institute of Technology,
WPAFB OH 45433-7765

**8. PERFORMING ORGANIZATION REPORT NUMBER**

AFIT/GSM/LAS/95S-3

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

ASC/QI
WPAFB OH 45433-7126

**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES**

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for public release; distribution unlimited

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

Measurement provides factual information which is necessary for effective control of business processes. Implementing a Total Quality Management (TQM) philosophy is a common process in many organizations today. The United States Air Force is using the Malcolm Baldrige National Quality Award criteria to measure organizational performance in implementing the Quality Air Force (QAF) initiative. Unfortunately, the Baldrige-based unit self-assessment (USA) process is an inconsistent measure due to its subjectivity, and is also time consuming to use.

This study determined that a new USA method based on the analytic hierarchy process (AHP) was not a significant improvement over the existing QAF-USA method. Specifically, this study developed a new USA scoring method by adapting the AHP to use existing QAF evaluation criteria. A group of 11 evaluators used this new AHP-USA method to score a portion of a USA report, and they also compared the AHP-USA method to the QAF-USA method to gauge its understandability, usability, believability and applicability. The resulting data was used to determine the overall feasibility and desirability of using the new method as a replacement for the QAF-USA method.

**14. SUBJECT TERMS**

Management, Total Quality Management (TQM), Management Planning and Control, Measure Theory, Decision Support Systems, Analytic Hierarchy Process (AHP), Malcolm Baldrige National Quality Award, Unit Self-assessment (USA)

**15. NUMBER OF PAGES** 171

**16. PRICE CODE**

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | UL |